Recent Advances and Opportunities in Adversarial Robustness

Jingfeng Zhang
SDU (2012 – 2016, Bachelor in CS)
NUS, Singapore (2016 – 2020, Ph.D.)
RIKEN-AIP, Tokyo (2021 – Present, Postdoctoral Researcher)
April 2021



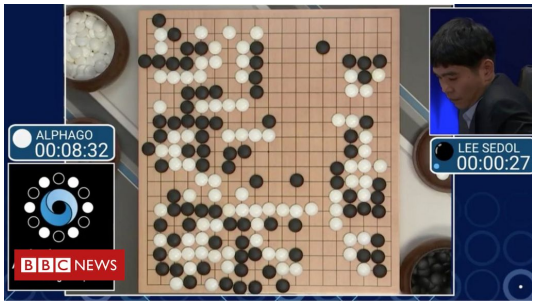
zjfheart

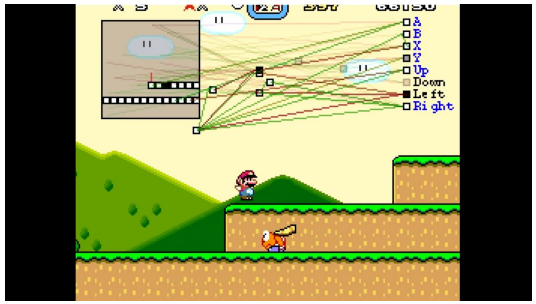

@zjf_heart

Artificial Intelligence (AI) exceeds human ability in many tasks

Image Classification

Reinforcement Learning

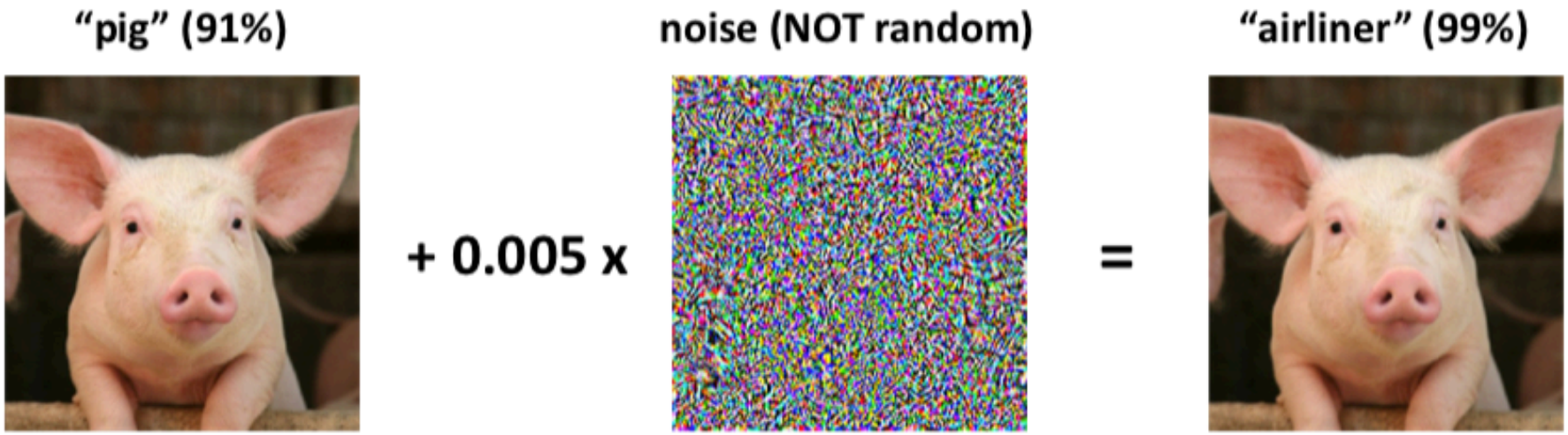




Natural language processing

Alexa, order me a large pizza!

Wait! The superiority of the current AI is just an illusion!



Natural data

Adversarial data

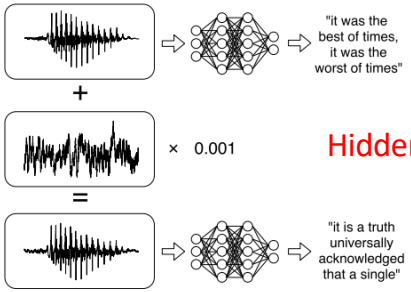
AI makes the low pig flying high!

The images & the joke come from Aleksander Madry’s group.

Motivation - Adversarial data pose threat to AI's deployment.

Glasses

[Sharif Bhagavatula Bauer Reiter 2016]

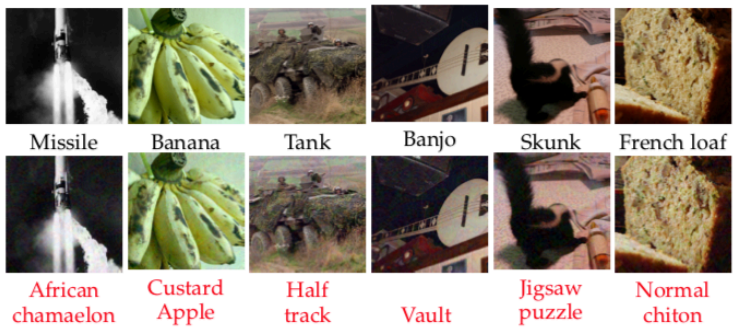


Hidden voice command

[Carlini Wagner 2018]

Small stickers

[Eykholt Evtimov Fernandes Li Rahmati Xiao Prakash Kohno Song 2018]



[Mopuri Ganeshan Babu 2018]

Fortunately, adversarial training is currently the most effective approach towards countering this threat!

THREAT!

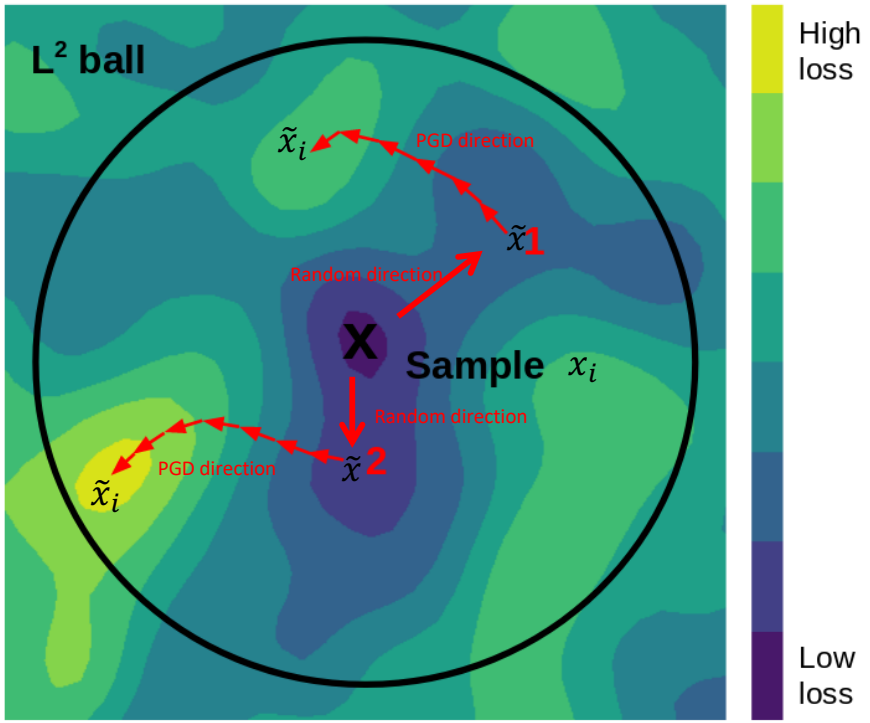
Preliminary – Adversarial data

Objective:

$$\tilde{x}_i = argmax_{\tilde{x} \in B(x_i)} \ell(f(\tilde{x}), y_i)$$

Find an adversarial data $\tilde{x}_i$ (within the norm ball $B(x_i)$) to maximize the loss $\ell(f(\tilde{x}), y_i)$; the norm ball constraint ensures the perseverance of the same semantic meaning of the adversarial data.

Images modified from https://towardsdatascience.com/know-your-enemy-7f7c5038bdf3



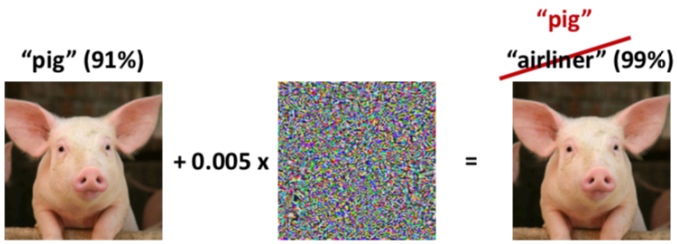
Method:

Projected gradient descent (PGD) –given a starting point $x^{(0)} \in \mathcal{X}$ and step size α, PGD works as followed:

$$x^{(t+1)} = \Pi_{B(x^{(0)})}\left(x^{(t)} + \alpha\, sign\left(\nabla_{x^{(t)}} \ell\left(f_\theta\left(x^{(t)}\right), y\right)\right)\right), t \in N$$

$\Pi_{B(x^{(0)})}$ projects adversarial data $x^{(t)}$ back onto the norm ball if $x^{(t)}$ exceeds the norm ball boundary; α is a small step size.

Preliminary – Standard adversarial training



Minimax formulation:

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(\tilde{x}_i), y_i), \text{ where } \tilde{x}_i = argmax_{x \in B(x_i)} \ell(f(\tilde{x}), y_i)$$

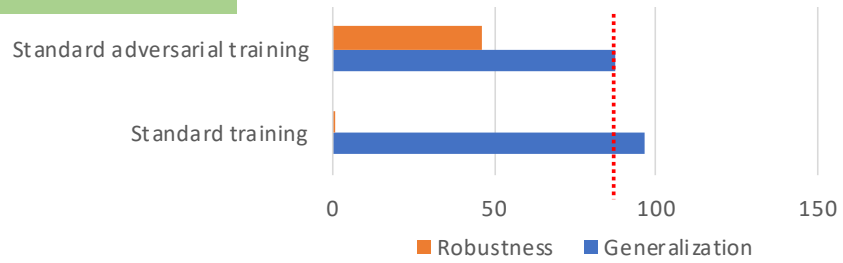
Outer minimization — Inner maximization

Realization:

[Madry Kakelov Schmidt Tsipras Vladu 2019]

Alternatively conduct steps (1) and (2):
(1) generate *most adversarial data* maximizing the loss (commonly using PGD method);
(2) minimize loss on the generated adversarial data w.r.t. model f parameters.

Empirical results:



In standard adversarial training,
although robustness gets improved, accuracy gets hurt.

Some studies argued "inherent tradeoff between robustness and accuracy"; is that true?

News!

Our work challenges two foundations of standard adversarial training.

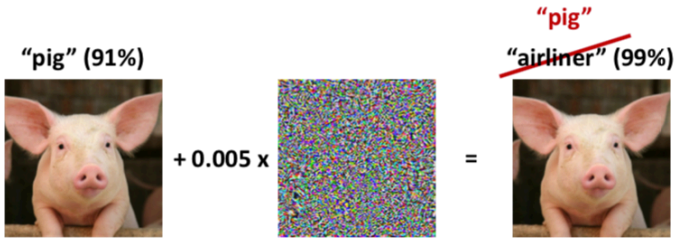
(1) the common belief that the minimax formulation is indispensable,

(2) the common belief of the inherent tradeoff between adversarial robustness and standard generalization.

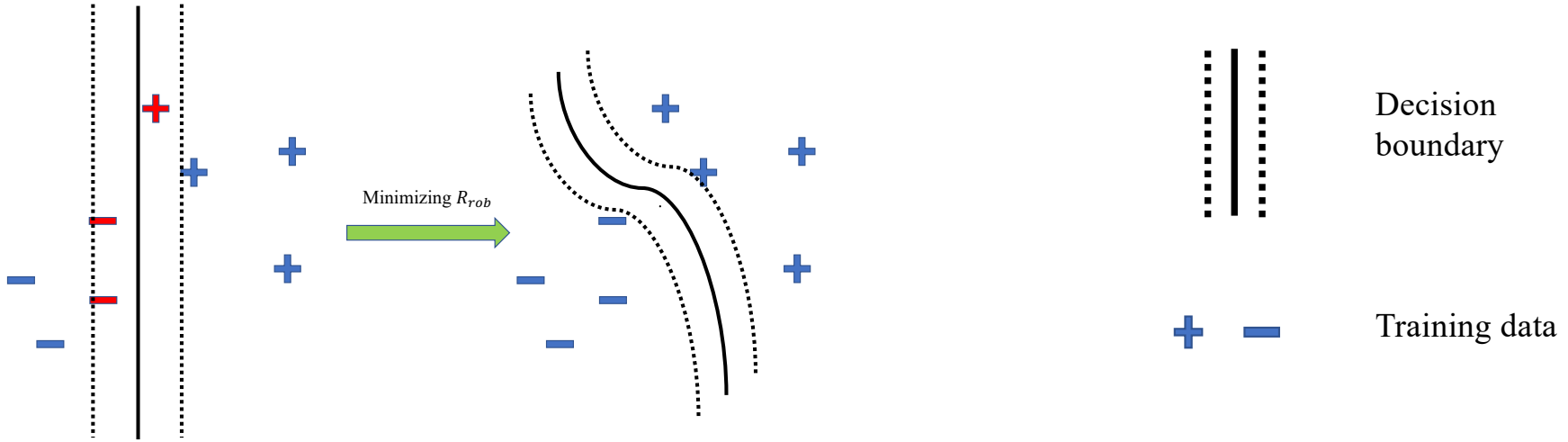
The following contents are based on the following papers:

- Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli, **Attacks Which Do Not Kill Training Make Adversarial Learning Stronger**, ICML 2020.
- Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli, **Geometry-aware Instance-reweighted Adversarial Training**, ICLR 2021 (Oral)

Purpose of adversarial learning



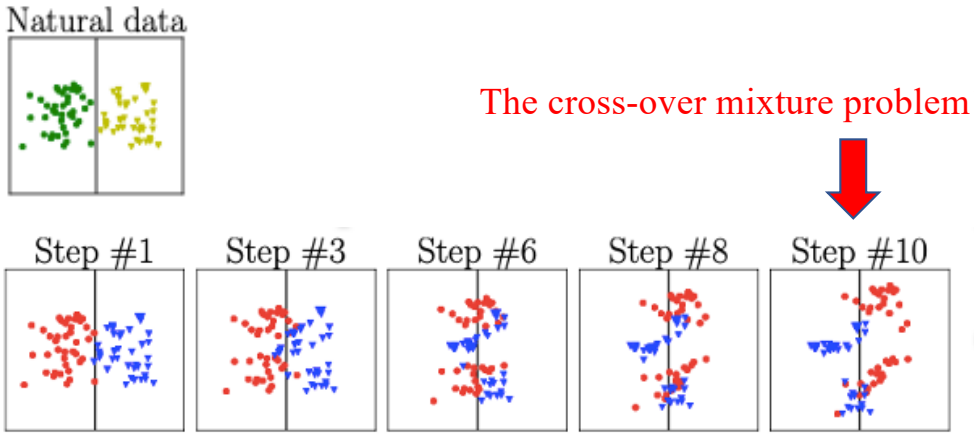
- **Adversarial data** can easily fool the standard trained classifier.
- **Adversarial training** so far is the most effective method for obtaining the adversarial robustness (against adversarial data) of the trained classifier.



Purpose 1: correctly classify the data.
Purpose 2: make the decision boundary thick so that no data is encouraged to fall inside the decision boundary.

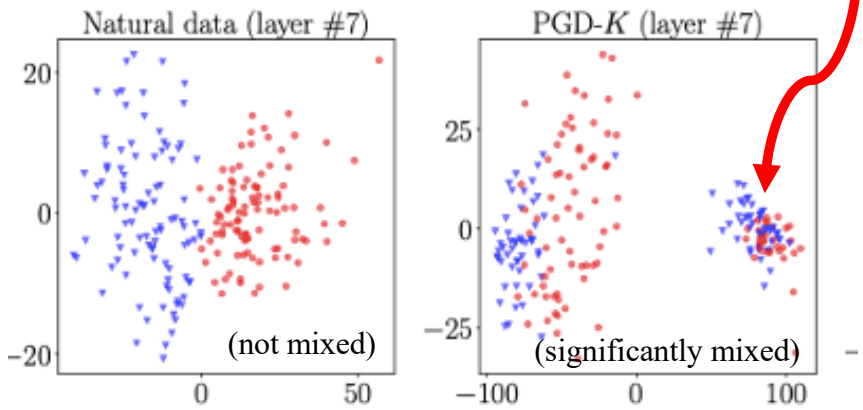
The minimax formulation is pessimistic.

- Many existing studies found the minimax-based adversarial training causes the severe degradation of the standard generalization. Why?



The adversarial data generated by PGD

Most adversarial data get significantly mixed



In the classification of the CIFAR-10 dataset, the cross-over mixture problem may not appear in the input space, but in the middle layers.

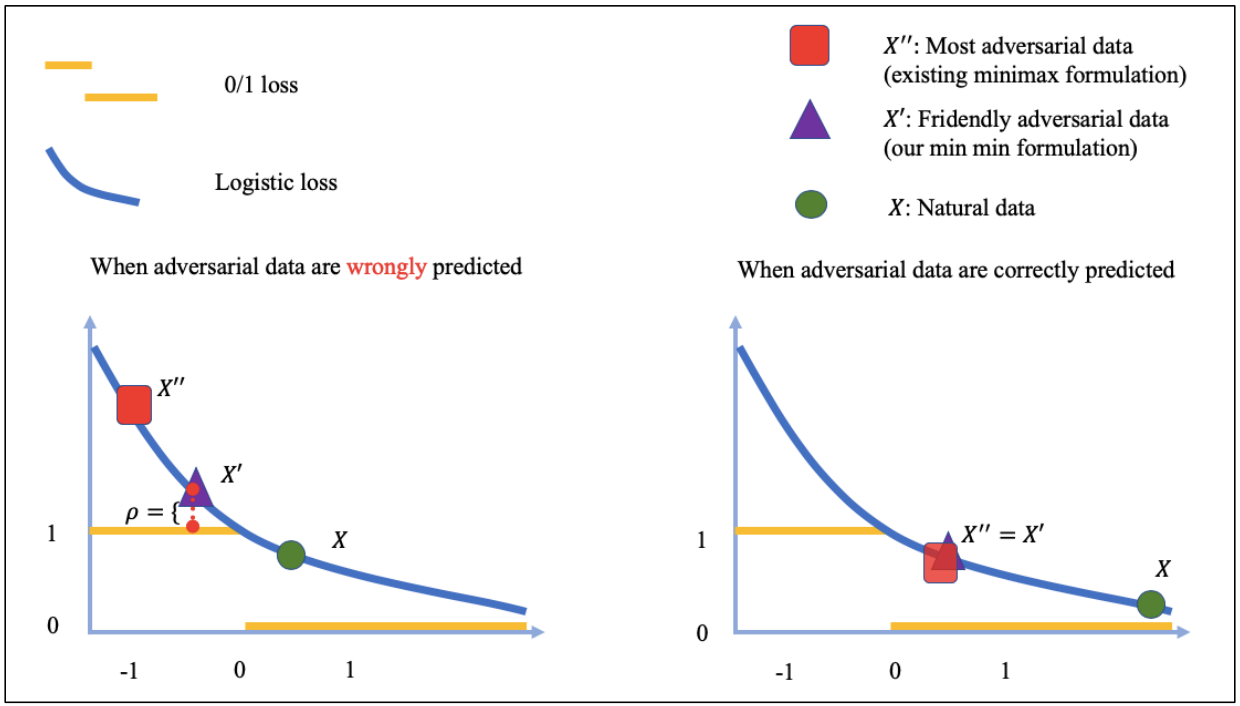
Is the minimax formulation suitable to the adversarial training?

Min-min formulation for the adversarial training

- The outer minimization keeps the same.
- Instead of generating *most adversarial data* $\tilde{x}_i$ via inner maximization, we generate *friendly adversarial data* $\tilde{x}_i$ as follows:

$$\tilde{x}_i = \arg min_{\tilde{x} \in B(x_i)} \ell(f(\tilde{x}), y_i) \text{ s.t. } \ell(f(\tilde{x}), y_i) - \min_{y \in \mathcal{Y}} \ell(f(\tilde{x}), y_i) \geq \rho$$

- The constraint firstly ensures $y_i \neq \arg\min_{y \in \mathcal{Y}} \ell(f(\tilde{x}), y_i)$ or $\tilde{x}$ is misclassified, and secondly ensures the wrong prediction of $\tilde{x}$ is better than the desired prediction y_i by at least the margin ρ in terms of the loss value.



Comparisons between minimax formulation and our proposed min-min formulation

Theoretical results: A tight upper bound on the adversarial risk

The adversarial risk $R_{rob}(f) := \mathbb{E}_{(X,Y\in D)}\mathbb{1}\{\exists\, X' \in B(X): f(X') \neq Y\}$

Zhang, Hongyang, et al. "Theoretically principled trade-off between robustness and accuracy." ICML 2019

Minimizing the adversarial risk captures the two purposes of the adversarial training:
(a) correctly classify the natural data and (b) make the decision boundary thick.

Theorem 1. For any classifier f, any non-negative surrogate loss function ℓ which upper bounds the 0/1 loss, and any probability distribution $\mathcal{D}$, we have

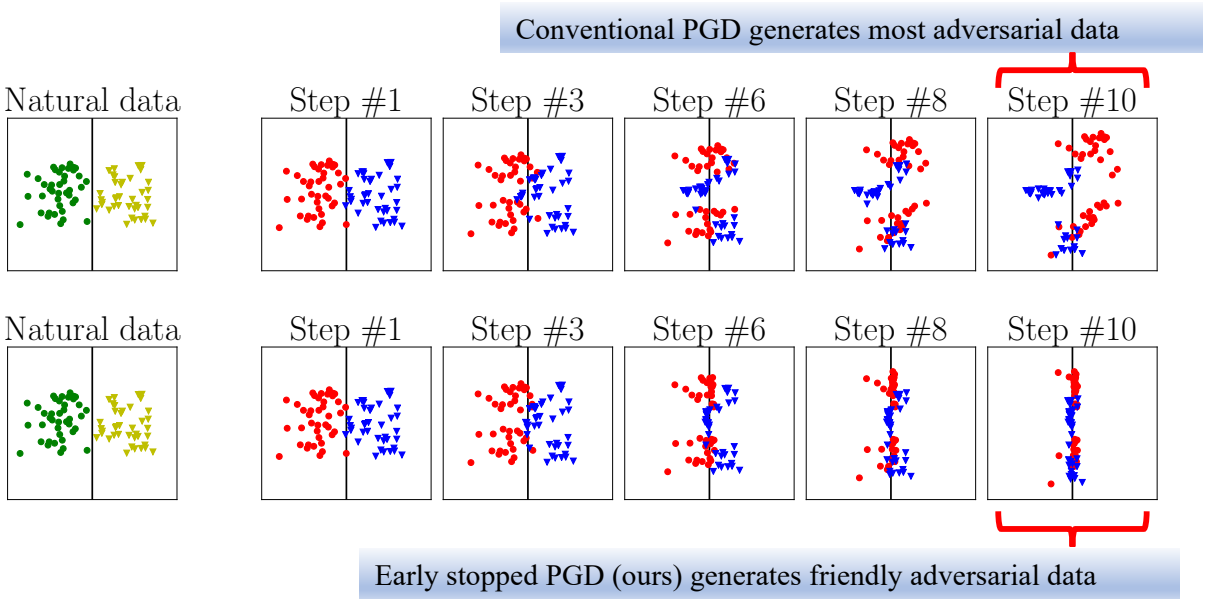
$$\mathcal{R}_{\text{rob}}(f) \leq \underbrace{\mathbb{E}_{(X,Y)\sim\mathcal{D}}\ell(f(X),Y)}_{\text{For standard test accuracy}} + \underbrace{\mathbb{E}_{(X,Y)\sim\mathcal{D},X'\in\mathcal{B}_\epsilon[X,\epsilon]}\ell^*(f(X'),Y)}_{\text{For robust test accuracy}},$$

where

$$\ell^* = \begin{cases} \min \ell(f(X'),Y) + \rho, & \text{if } f(X') \neq Y, \\ \max \ell(f(X'),Y), & \text{if } f(X') = Y. \end{cases}$$

Theoretically, our min-min formulation facilitates a tighter upper bound on the adversarial risk, compared with minmax formulation.

Realization of our min-min formulation – friendly adversarial training (FAT)



Algorithm 1 Early stopped PGD-K-τ

Input: data $x \in \mathcal{X}$, label $y \in \mathcal{Y}$, model f, loss function ℓ, maximum PGD step K, step τ, perturbation bound ϵ, step size α

Output: $\tilde{x}$

$\tilde{x} \leftarrow x$

while $K > 0$ **do**

  if $\arg\max_i f(\tilde{x}) \neq y$ and $\tau = 0$ **then**

    break

  else if $\arg\max_i f(\tilde{x}) \neq y$ **then**

    $\tau \leftarrow \tau - 1$

  end if

  $\tilde{x} \leftarrow \Pi_{\mathcal{B}[x,\epsilon]}\left(\alpha\,\mathrm{sign}(\nabla_{\tilde{x}}\ell(f(\tilde{x}), y)) + \tilde{x}\right)$

  $K \leftarrow K - 1$

end while

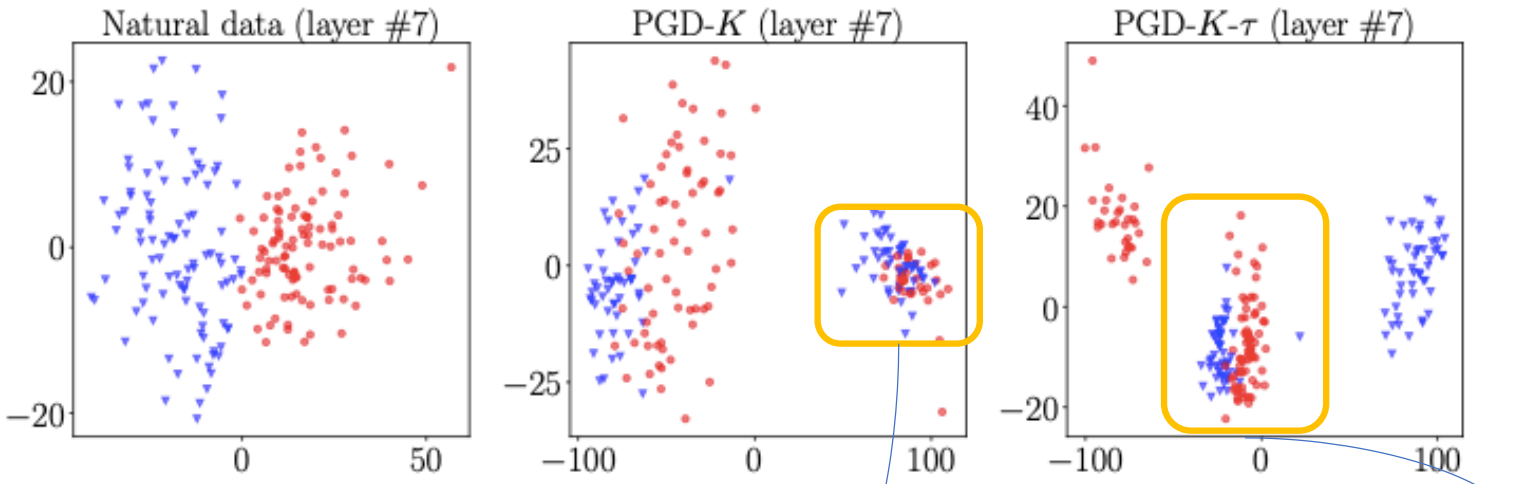
K is the maximum allowed PGD step numbers; step τ controls number of the extra steps once a misclassified adversarial data is found.

When τ = K, it recovers the conventional PGD method.

For updating the model, friendly adversarial training (FAT) employs the **friendly adversarial data** generated by **early stopped PGD**.

Benefits (a): Alleviate the cross-over mixture problem

- In the classification of the CIFAR-10 dataset, the cross-over mixture problem may not appear in the input space, but in the middle layers.

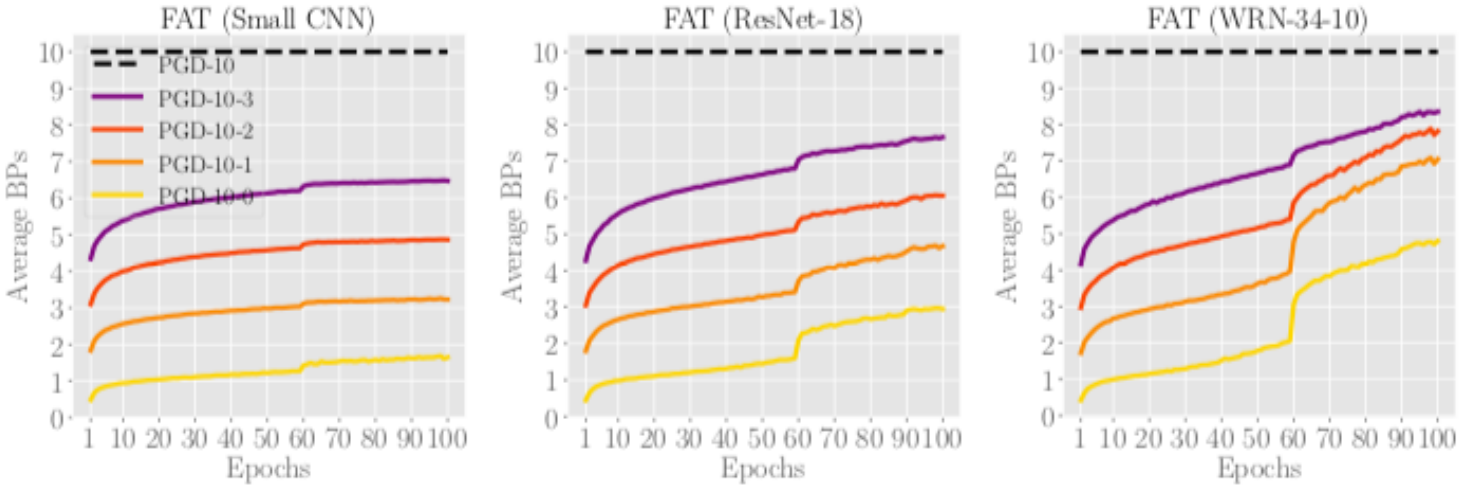


Natural data (not mixed)

Most adversarial data generated by conventional PGD (significantly mixed)

Friendly adversarial data generated by early stopped PGD (not significantly mixed)

Benefits (b): FAT is computationally efficient.



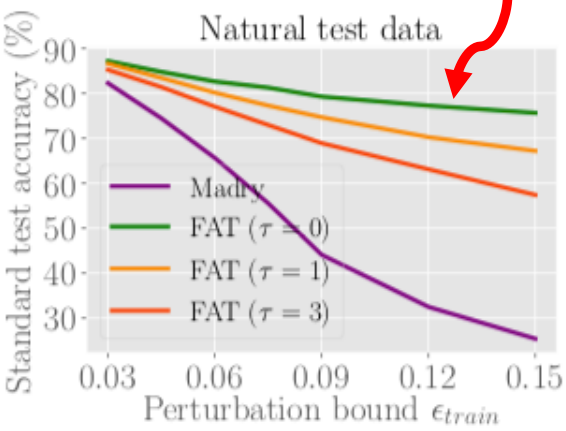
We report the average backward propagations (BPs) per epoch over training process.

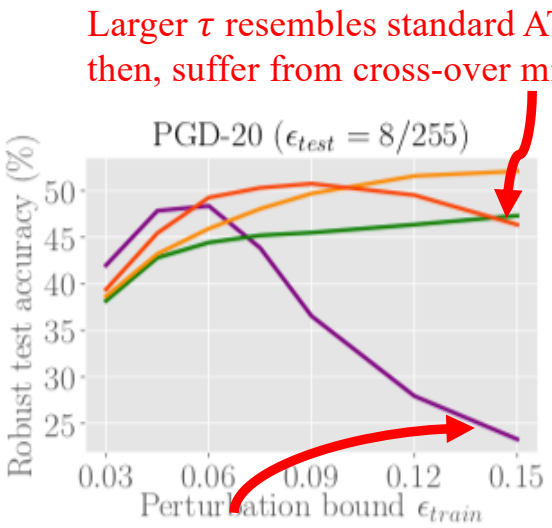
Dashed line (PGD-10) is existing standard adversarial training based on conventional PGD.
(Note that PGD-10 equals to PGD-10-10 with $\tau = 10$.)

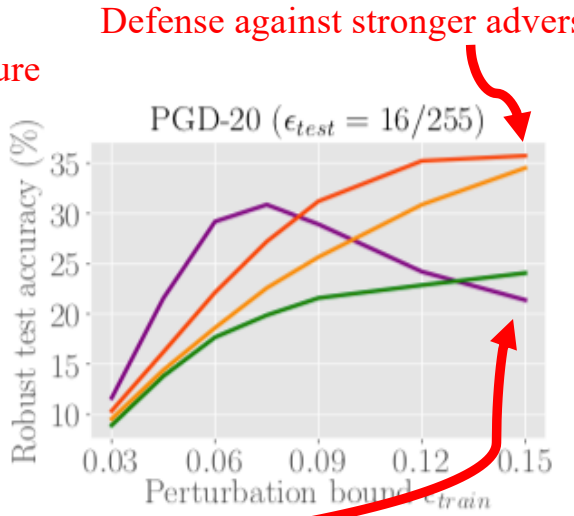
Solid lines are friendly adversarial trainings based on early stopped PGD with different τ.

Compared with standard AT (PGD-10), the dashed line minus solid lines are the saved BPs by FAT.

Benefits (c): FAT can enable larger defense parameter ϵ_{train}







For CIFAR-10 dataset, we adversarially train deep neural networks with $\epsilon_{train} \in [0.03, 0.15]$, and evaluate each robust model with 3 evaluation metrics (1 natural generalization metric + 2 robustness metrics with different adversary strength ϵ_{test}.)

The purple line represents standard adversarial training (Madry's; PGD-10-10).

The red, orange and green lines represent our friendly adversarial training with different configurations τ (PGD-10-τ).

Benefits (d): Benchmarking on Wide ResNet.

Table 1. Evaluations (test accuracy) of deep models (WRN-32-10) on CIFAR-10 dataset

Defense	Natural	FGSM	PGD-20	C&W$_\infty$	PGD-100
Madry	87.30	56.10	45.80	46.80	-
CAT	77.43	57.17	46.06	42.28	-
DAT	85.03	63.53	48.70	47.27	-
FAT ($\epsilon_{train} = 8/255$)	**89.34** $\pm$ 0.221	65.52 $\pm$ 0.355	46.13 $\pm$ 0.409	46.82 $\pm$ 0.517	45.31 $\pm$ 0.531
FAT ($\epsilon_{train} = 16/255$)	87.00 $\pm$ 0.203	**65.94** $\pm$ 0.244	**49.86** $\pm$ 0.328	**48.65** $\pm$ 0.176	**49.56** $\pm$ 0.255

Results of Madry, CAT and DAT are reported in (Wang et al., 2019). FAT has the same evaluations.

Wang, Yisen, et al. "On the convergence and robustness of adversarial training." ICML 2019

Table 2. Evaluations (test accuracy) of deep models (WRN-34-10) on CIFAR-10 dataset

Defense	Natural	FGSM	PGD-20	C&W$_\infty$	PGD-100
TRADES ($\beta = 1.0$)	88.64	56.38	49.14	-	-
FAT for TRADES ($\epsilon_{train} = 8/255$)	**89.94** $\pm$ 0.303	**61.00** $\pm$ 0.418	**49.70** $\pm$ 0.653	49.35 $\pm$ 0.363	48.35 $\pm$ 0.240
TRADES ($\beta = 6.0$)	84.92	61.06	56.61	**54.47**	55.47
FAT for TRADES ($\epsilon_{train} = 8/255$)	**86.60** $\pm$ 0.548	**61.97** $\pm$ 0.570	55.98 $\pm$ 0.209	54.29 $\pm$ 0.173	55.34 $\pm$ 0.291
FAT for TRADES ($\epsilon_{train} = 16/255$)	84.39 $\pm$ 0.030	61.73 $\pm$ 0.131	**57.12** $\pm$ 0.233	54.36 $\pm$ 0.177	**56.07** $\pm$ 0.155

Results of TRADES ($\beta = 1.0$ and 6.0) are reported in (Zhang et al., 2019b). FAT for TRADES has the same evaluations.

Zhang, Hongyang, et al. "Theoretically principled trade-off between robustness and accuracy." ICML 2019

FAT can improve standard test accuracy while maintaining the superior adversarial robustness.

This challenges the inherent tradeoff between robustness and accuracy!

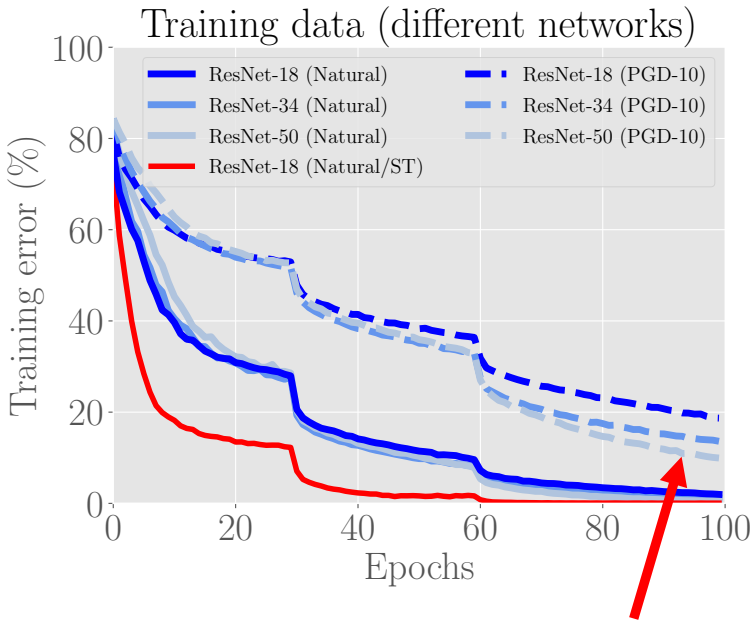
How about other direction?

- The other direction---whether we can improve the adversarial robustness while retaining the standard accuracy---is conceptually and practically more interesting.

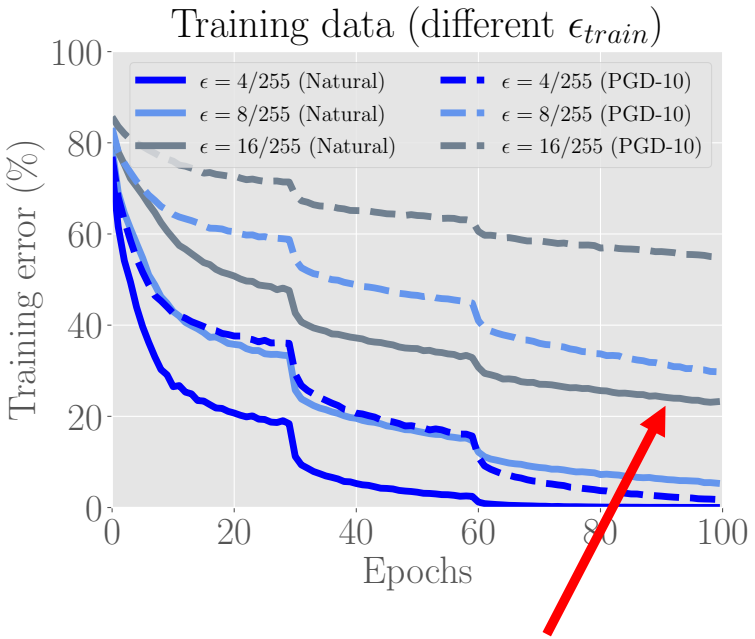
Next, we are going to show this direction is also achievable!

Fact 1: model capacity is often insufficient in adversarial training.

- This point is very counter-intuitive in deep learning.



The networks hardly reach zero error on the adversarial training data.



A slightly larger perturbation bound ϵ_{train} significantly uncovers this insufficiency of the model capacity

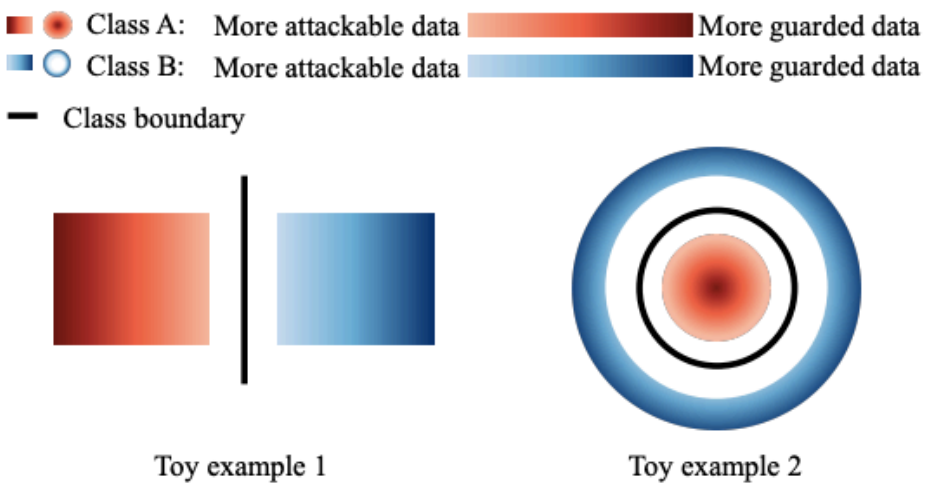
Adversarial training (AT) on CIFAR-10 dataset.

AT has very strong smoothing effect!

Smooth large neighborhood

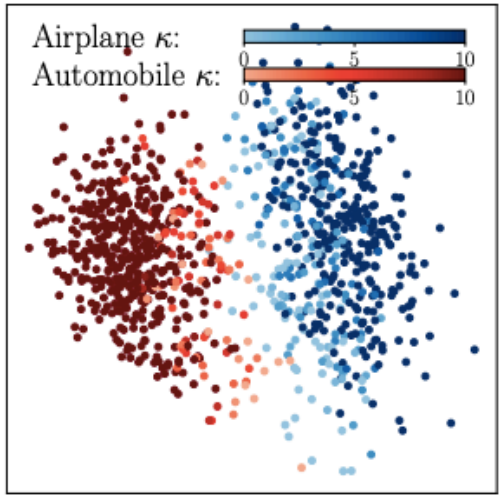
$|1+\epsilon|^{\text{input\_dim}}$

Fact 2: natural data points have different degrees of robustness.



Toy example 1

Toy example 2

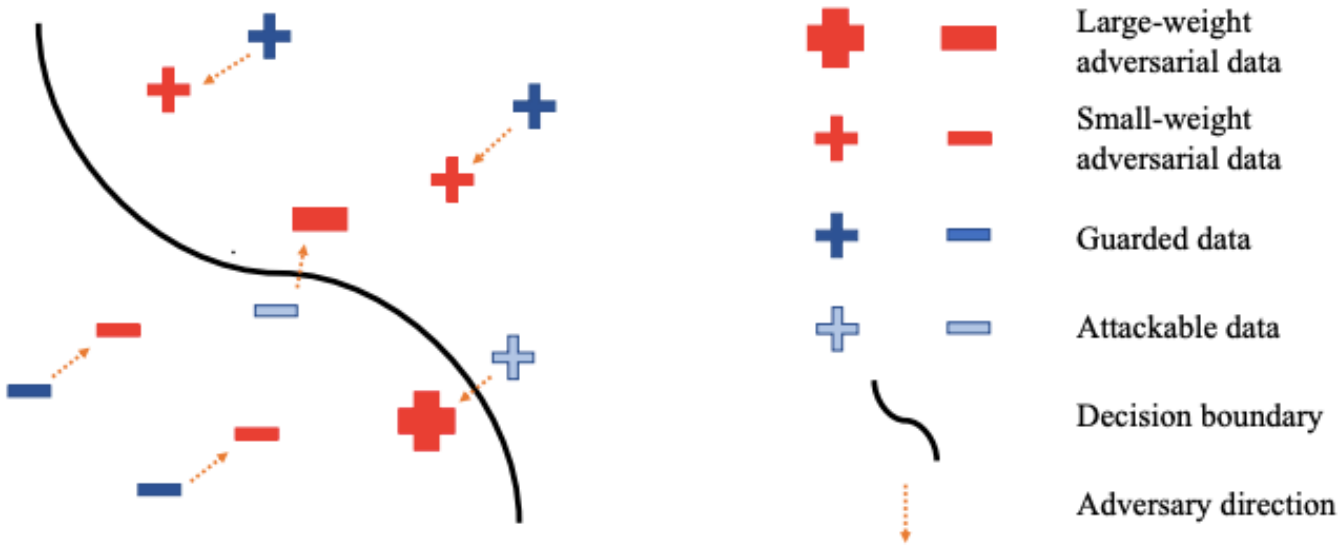


The model's output distribution of two randomly selected classes from the CIFAR-10 dataset.

More attackable/guarded data are closer to/farther away from the class boundary.

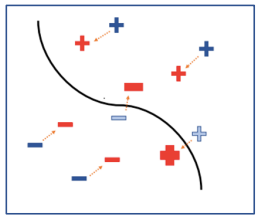
Given limited model capacity, the data's adversarial variants should have unequal importance for fine-tuning the decision boundary that approximates the class boundary.

Geometry-Aware Instance-Reweighted Adversarial Training (GAIRAT)



For updating the model, GAIRAT explicitly gives larger weights on the losses of adversarial data (larger red), whose natural counterparts are closer to the decision boundary (lighter blue).

Realization of GAIRAT

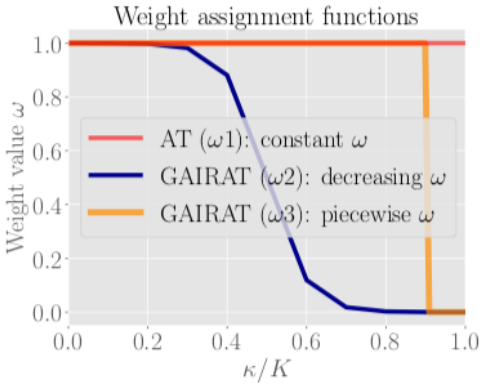


- $\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \omega(x_i, y_i) \ell(f(\tilde{x}_i), y_i)$,

where $\tilde{x}_i$ is most/friendly adversarial data, x_i is natural data;$\omega(x_i, y_i)$ is instance-dependent weight assignment function, whose values are based on the geometry distance of natural data (x_i, y_i) from the decision boundary.

- How to approximate the geometry distance of data (x_i, y_i)?

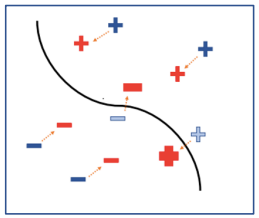
Our solution: the least number of PGD iteration κ (PGD step number) that the PGD method requires to generate a misclassified adversarial variant, given maximum allowed PGD steps K.



We call κ the data's geometry value.

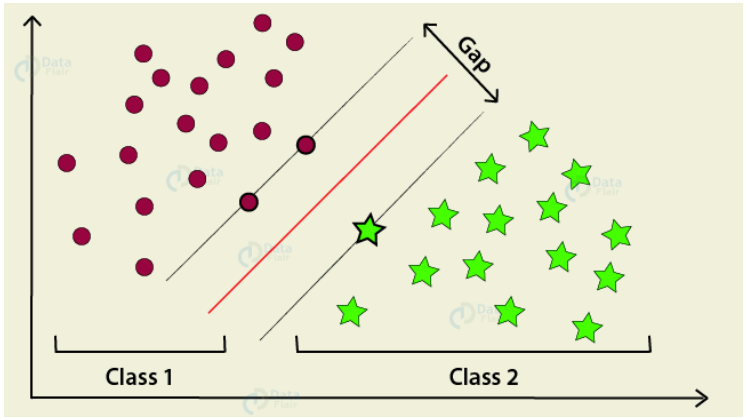
guarded data have larger κ,
attackable data have smaller κ.

$\omega(x_i, y_i)$ assigns weights inversely proportional to $\kappa(x_i, y_i)/K$.

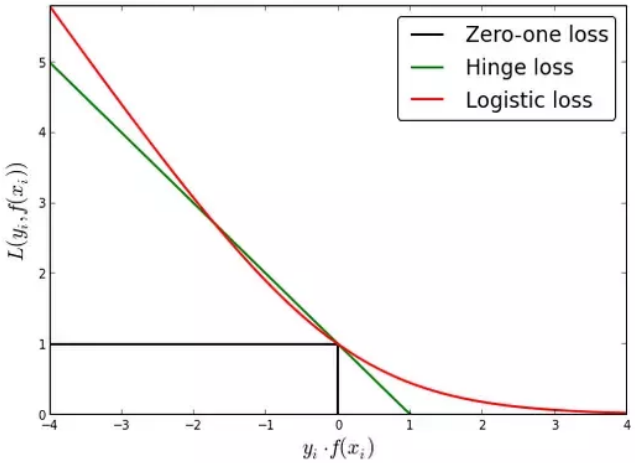


GAIRAT's relationship with traditional machine learning methods - SVM

https://data-flair.training/blogs/svm-support-vector-machine-tutorial/



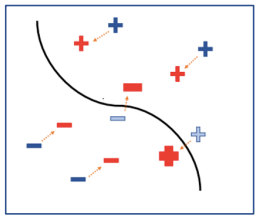
Support vector machine (SVM)



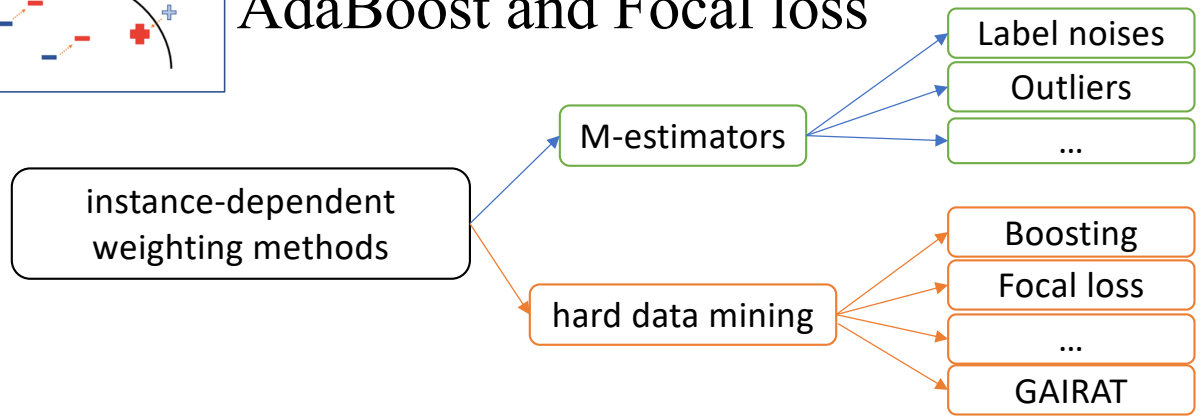
Hinge loss for training SVM

In standard training, the magnitude of loss can naturally capture the data's geometry distance to decision boundary.

But in adversarial training, there is a blocking effect.
The magnitude of loss on adversarial data fails to do so because adversarial data are generated to maximize the loss! Therefore, we need GAIRAT!



GAIRAT's relationship with traditional machine learning methods – AdaBoost and Focal loss



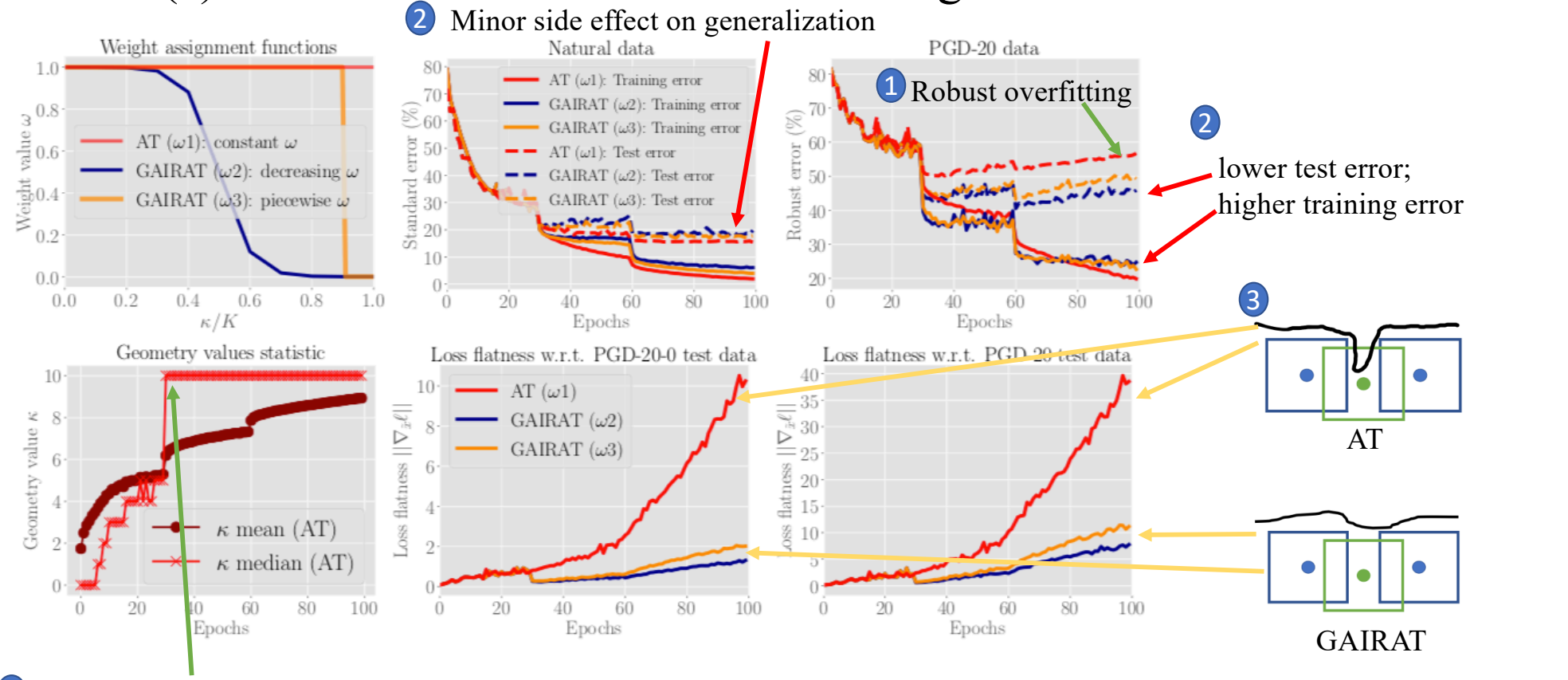
- Boosting algorithms such as AdaBoost select harder examples to train subsequent classifiers.
- Focal loss is specially designed loss function for mining hard data and misclassified data.
- Boosting and Focal loss leverage the data's losses for measuring the data's hardness;
- by comparison, our GAIRAT measures the hardness by how difficulty the natural data are attacked (i.e., geometry value κ). This is a new measurement.

GAIRAT's relationship with geometric studies of DNN.

- Researchers in adversarial robustness employed the first-order or second-order derivatives w.r.t. input data to explore the DNN's geometric properties.

- Instead, we have a complementary but different argument: data points themselves are geometrically different regardless of DNN.
- The geometry value κ in adversarial training (AT) is an approximated measurement of data's geometric properties due to the AT's smoothing effect.

Benefits (a): GAIRAT relieves robust overfitting



1 Standard AT engenders large number of guarded data, overwhelming the small number of attackable data, which leads to robust overfitting (red lines).

GAIRAT - small weight - large number
GAIRAT - large weight - small number

Benefits (b): Benchmarking on Wide ResNet.

Table 1: Test accuracy of WRN-32-10 on CIFAR-10 dataset

Defense	Best checkpoint						Last checkpoint					
	Natural	Diff.	PGD-20	Diff.	PGD+	Diff.	Natural	Diff.	PGD-20	Diff.	PGD+	Diff.
AT	86.92 ± 0.24	-	51.96 ± 0.21	-	51.28 ± 0.23	-	86.62 ± 0.22	-	46.73 ± 0.08	-	46.08 ± 0.07	-
FAT	$\mathbf{89.16 \pm 0.15}$	+2.24	51.24 ± 0.14	−0.72	46.14 ± 0.19	−5.14	88.18 ± 0.19	+1.56	46.79 ± 0.34	+0.06	45.80 ± 0.16	−0.28
GAIRAT	85.75 ± 0.23	−1.17	$\mathbf{57.81 \pm 0.54}$	+5.85	$\mathbf{55.61 \pm 0.61}$	+4.33	85.49 ± 0.25	−1.13	$\mathbf{53.76 \pm 0.49}$	+7.03	$\mathbf{50.32 \pm 0.48}$	+4.24
GAIR-FAT	88.59 ± 0.12	+1.67	56.21 ± 0.52	+4.25	53.50 ± 0.60	+2.22	$\mathbf{88.44 \pm 0.10}$	+1.82	50.64 ± 0.56	+3.91	47.51 ± 0.51	+1.43

GAIRAT can improve the adversarial robustness, while maintaining the standard accuracy. The other direction is achieved!

Combining two directions (FAT + GAIRAT), i.e., GAIR-FAT, we can improve both robustness and accuracy of standard AT

Benefits (c): GAIRAT can obtain the competitive results!

- We incorporate 500,000 auxiliary CIFAR-10 data + 50,000 CIFAR-10 training data.

*500,000 auxiliary CIFAR-10 data by **[Carmon Raghunathan Schmidt Liang Duchi, 2019]**

- Our "geometry aware instance reweighted" method (using standard WRN-28-10) achieves the standard test accuracy (89.8%) and the robustness score (60.9%) using a subset of AA attack (1/5 of test data).

```
initial accuracy: 89.80%
apgd-ce - 1/2 - 133 out of 500 successfully perturbed
apgd-ce - 2/2 - 88 out of 398 successfully perturbed
robust accuracy after APGD-CE: 67.70% (total time 204.3 s)
apgd-t - 1/2 - 54 out of 500 successfully perturbed
apgd-t - 2/2 - 14 out of 177 successfully perturbed
robust accuracy after APGD-T: 60.90% (total time 1327.4 s)
fab-t - 1/2 - 0 out of 500 successfully perturbed
fab-t - 2/2 - 0 out of 109 successfully perturbed
robust accuracy after FAB-T: 60.90% (total time 3014.6 s)
square - 1/2 - 0 out of 500 successfully perturbed
square - 2/2 - 0 out of 109 successfully perturbed
robust accuracy after SQUARE: 60.90% (total time 5033.1 s)
max Linf perturbation: 0.03100, nan in tensor: 0, max: 1.00000, min: 0.00000
robust accuracy: 60.90%
```

AA attack results

Summaries.

- Our contributions to the state of knowledge are:
- (a) We show that the minimax formulation is NOT indispensable for adversarial training; instead, we propose our min-min formulation, which can inspire more effective adversarial training methods.
- (b) We show that, in adversarial training, model capacity is not enough due to its over-smoothing effect.
- (c) We show that, given the limited model capacity, we should explicitly treat adversarial data differently.
- (d) We propose two effective strategies, i.e., FAT and GAIRAT, which challenge the common belief of the inherent trade-off.

Short introductions of the new works.

(1) Improve adversarial robustness further by exploring novel network structures---diverse-structured network (DS-Net).

Fact 1: the optimal network architectures in standard training (ST) would be no longer optimal in adversarial training (AT).

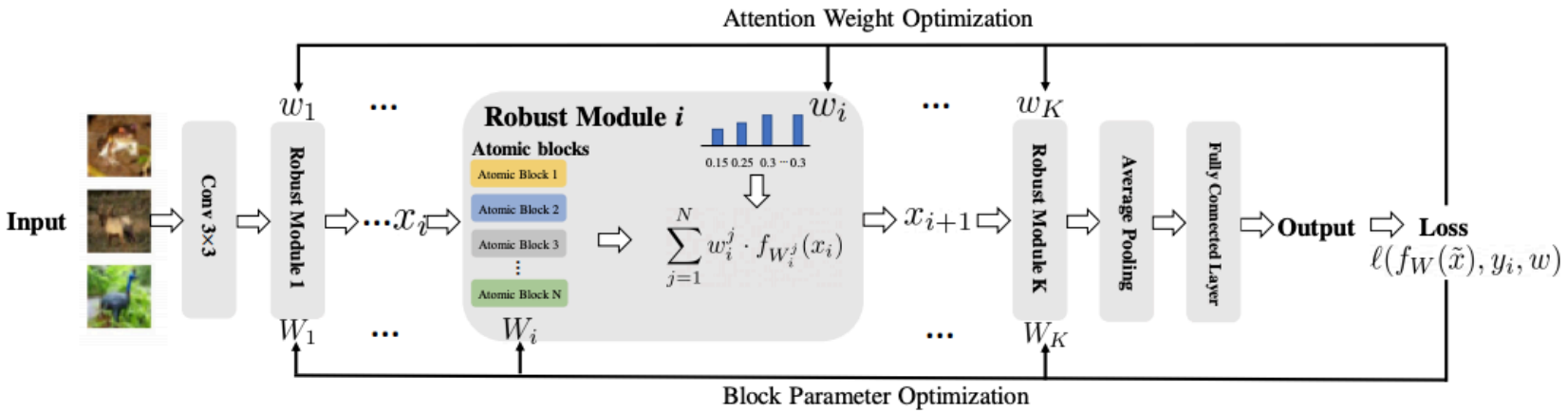
Model	Standard Acc.	Ranking	Robustness	Ranking
WRN-28-10	0.9646	1	0.4872	3
ResNet-62	0.9596	2	0.4855	4
DenseNet-121	0.9504	3	0.4993	2
MobileNetV2	0.9443	4	0.4732	6
AdaRKNet-62	0.9403	5	0.5016	1
ResNet-50	0.9362	6	0.4807	5

Fact 2: AT is time-consuming itself; if we directly search network architectures in AT over large search spaces (e.g., using NAS), the computation will be practically infeasible.

Reference: Learning Diverse-structured Networks for Adversarial Robustness, 2021

Our solution

- we propose a diverse-structured network (DS-Net), to significantly reduce the size of the search space:
- Instead of low-level operations (of NAS), we only consider predefined atomic blocks, where an atomic block is a time-tested building block such as the residual block, dense block and so on.
- Our novel network design strategy is a trade-off between exploring diverse structures and exploiting the best structures.

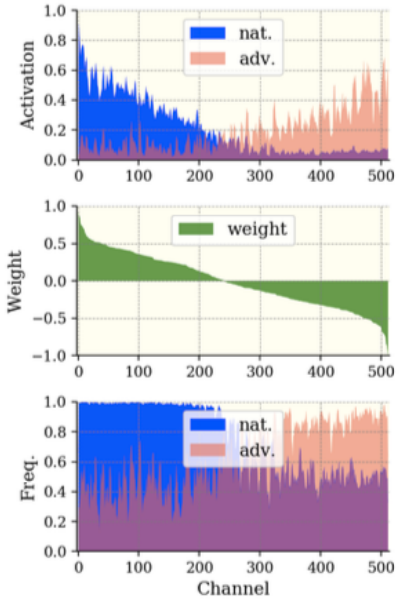


Reference: Learning Diverse-structured Networks for Adversarial Robustness, 2021

(2) Improve adversarial robustness further by exploring novel network structures---Channel-wise Importance-based Feature Selection (CIFS).

- The idea basic: aligning channels' activations of adversarial data with that of natural data.

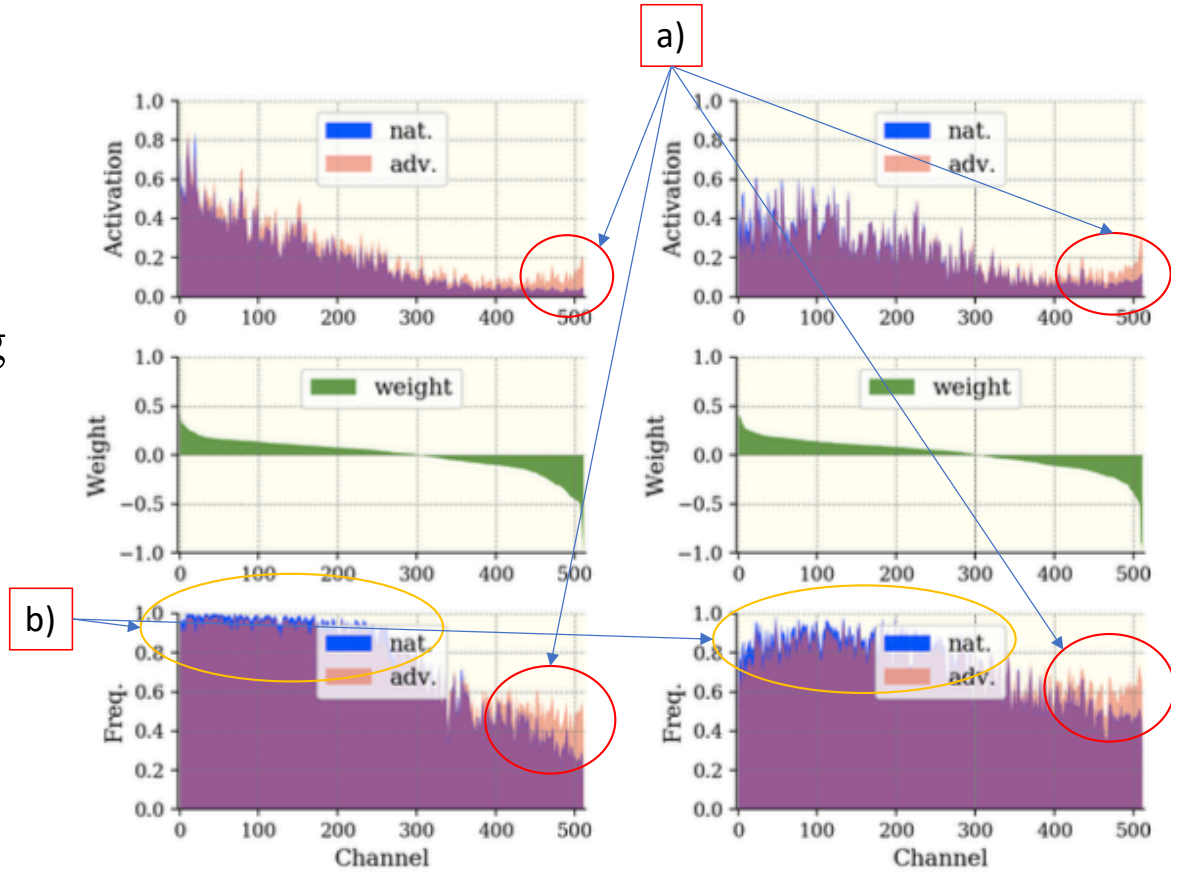
- Issue of standard training:



(b) Normal / "cat"

Reference:CIFS: Improving Adversarial Robustness of CNNs via Channel-wise Importance-based Feature Selection , 2021

- AT can largely alleviate this misalignment, but still has two pitfalls:
- a) The channels that are *negatively-relevant* (NR) (correspond to negative weight) to predictions are still over-activated when processing adversarial data.
- b) AT does not result in similar robustness for all classes: For the robust classes, channels with larger activation magnitudes are usually more *positively-relevant* (PR) to predictions.



(c) Adv. / "automobile"

69.0% robust acc. for the "automobile" class

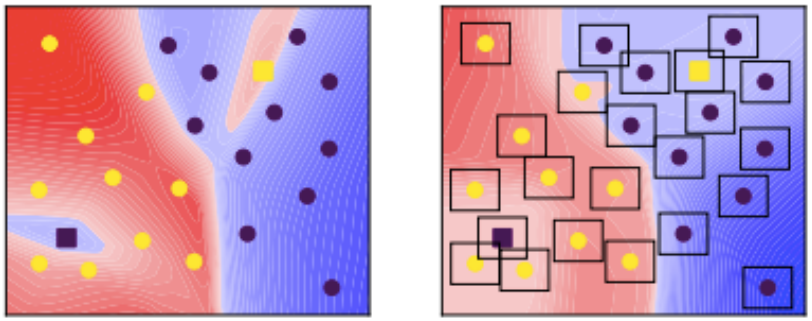
(d) Adv. / "cat"

16.7% robust acc. for the "cat" class

Reference:CIFS: Improving Adversarial Robustness of CNNs via Channel-wise Importance-based Feature Selection , 2021

(3) Adversarial training for other domains---label noises

Comparisons between standard training (ST)
and adversarial training (AT)

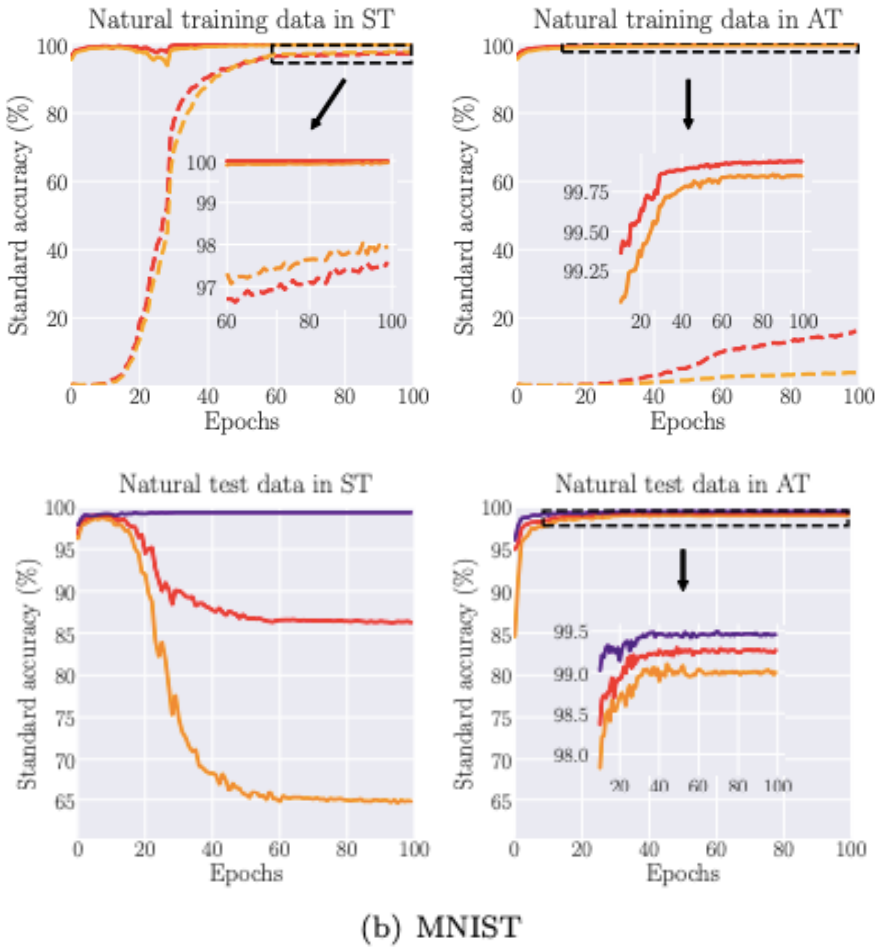


ST AT

AT has smoothing effect, naturally mitigate
the negative effects of label noises

Reference: Understanding the Interaction of Adversarial Training with Noisy Labels, 2021

What does this smoothing effect imply?



(b) MNIST

Different color means different levels of label noises

1) Mitigate label noises.

Solid lines denote the accuracy of correct training data, while dashed lines correspond to that of incorrect training data.

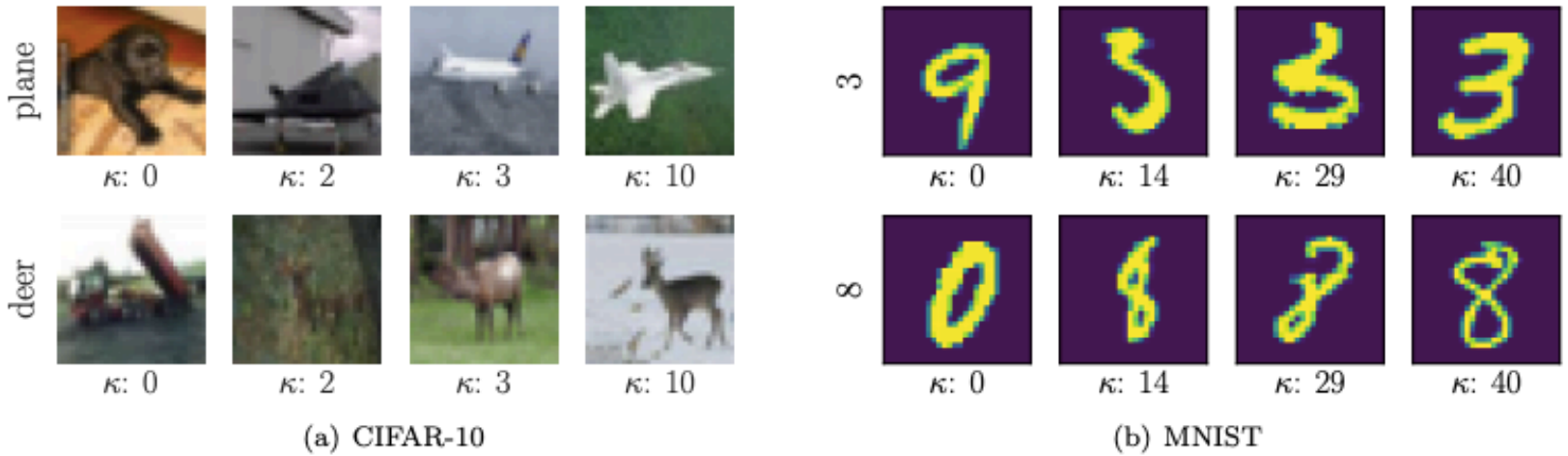
2) Stop generalization degradation

Solid lines denote the accuracy of correct test data.

Reference: Understanding the Interaction of Adversarial Training with Noisy Labels, 2021

What does this smoothing effect imply?

- Geometry value κ (PGD steps to find a misclassified data) is useful metric to distinguish noise-labeled data, rare data and classic data in AT.



Reference: Understanding the Interaction of Adversarial Training with Noisy Labels, 2021

(4) Two sample test is aware of adversarial data!

- Existing studies have showed that statistic tests cannot detect adversarial data. Carlini, N. and Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods.
- However, we ask ourselves: Are *natural data and adversarial data really from the same distributions?* (No! Then, where is the gap?)
- …
- We proposed *semantic-aware MMD* (SAMMD) test, which is indeed aware of adversarial attacks.

Reference: Maximum Mean Discrepancy is Aware of Adversarial Attacks, 2021

Challenges & opportunities: What is the next?

- 1) Design proper network structures for AT or input smoothing.
- 2) Build the comprehensive understanding of AT's traits such as robust overfitting, smoothing effect and so on.
- 3) Leverage AT techniques for other domains, e.g., pretraining, few shot learning, semi-supervised learning, label noises, and so on.
- 4) Design effective AT strategies on improving generalization and robustness.
- 5) Develop a suitable optimizer for AT.

Want to conduct research in the topic adversarial robustness?

ICML, NeurIPS, ICLR?

Welcome to talk with me.

Wechat: zjfheart
Email: jingfeng.zhang@riken.jp

We provide collaborated projects or numerous internship opportunities!

Locations: Tokyo, Shanghai, Hangzhou, Shenzhen, Hong Kong or online!

Thanks