SmoothMix: Training Confidence-calibrated Smoothed Classifiers for Certified Robustness

Jongheon Jeong[1], Sejun Park[2], Minkyu Kim[1], Heung-Chang Lee[3], Doguk Kim[4], Jinwoo Shin[1]

[1]KAIST [2]Vector Institute [3]Kakao Enterprise [4]Inha University

NeurIPS 2021

Background: Adversarial Examples

Deep neural networks (DNNs) are susceptible to adversarial noises δ

Stop Sign

$f(x)$

+

Noise

δ

=

Max Speed 100

$f(x+\delta)$

Fundamental question: Can we build DNNs that are robust to such noises?

$$f(x) = f(x+\delta), \quad \forall\delta : ||\delta||_2 < \epsilon$$

a classifier

The hardest part

Image source: https://deep.ghost.io/robust-attribution/

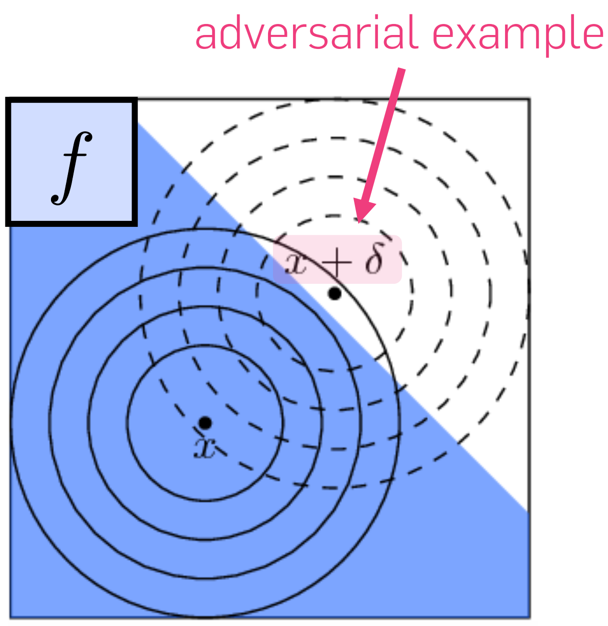
Background: Adversarial Training

Challenge: DNNs are too complex to regularize every $f(x+\delta)$

- Adversarial training (AT) [Madry et al., 2018]?

$$\min_{f} \mathbb{E}_{(x,y)} \left[\max_{\delta} \mathcal{L}(x+\delta, y; f) \right]$$

adversarial example



  - Only gives an empirical robustness
    - It is hard to guarantee that an AT-model is "indeed" robust
  - Harder to optimize and generalize [Schmidt et al., 2018]
  - Seems to require much larger network
    - AT does not saturate even at ResNet-638 on ImageNet [Xie & Yuille, 2020]

[Cohen et al., 2019] Certified adversarial robustness via randomized smoothing. ICML 2019.
[Schmidt et al., 2018] Adversarially Robust Generalization Requires More Data, NeurIPS 2018.
[Madry et al., 2018] Towards deep learning models resistant to adversarial attacks, ICLR 2018.
[Xie & Yuille, 2020] Intriguing properties of adversarial training at scale, ICLR 2020.

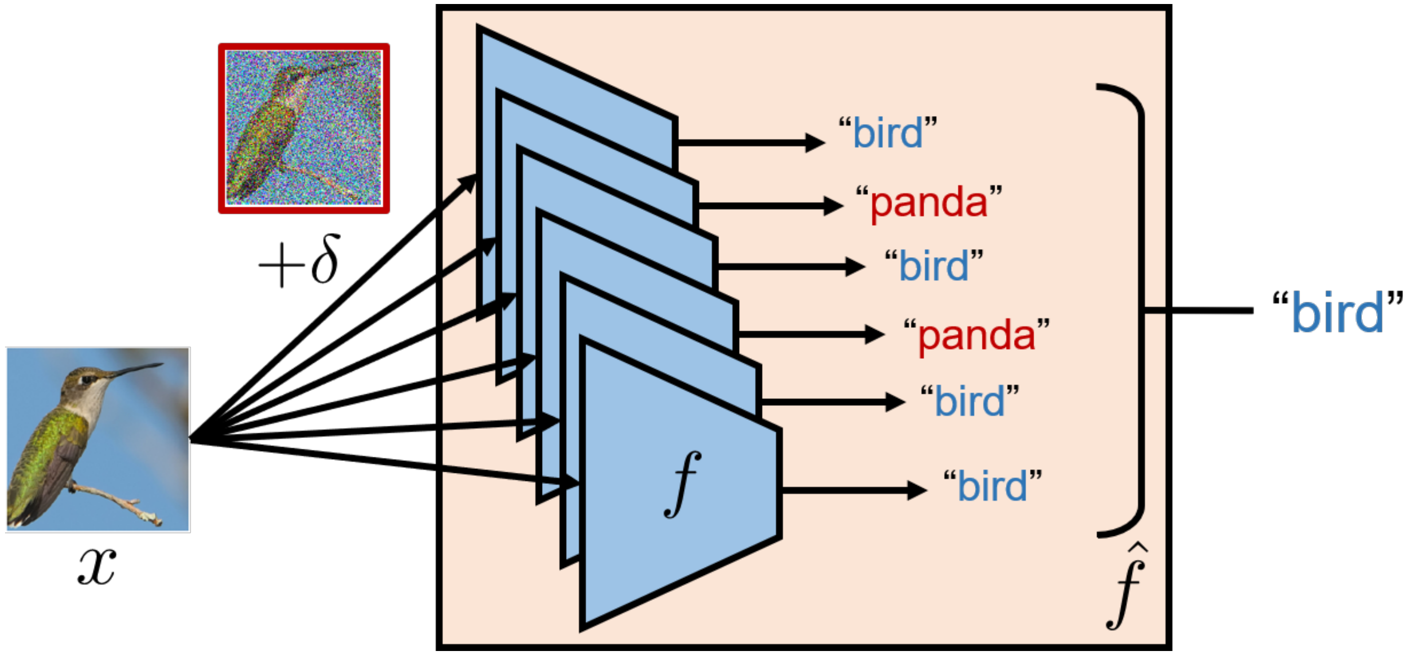
Background: Randomized Smoothing

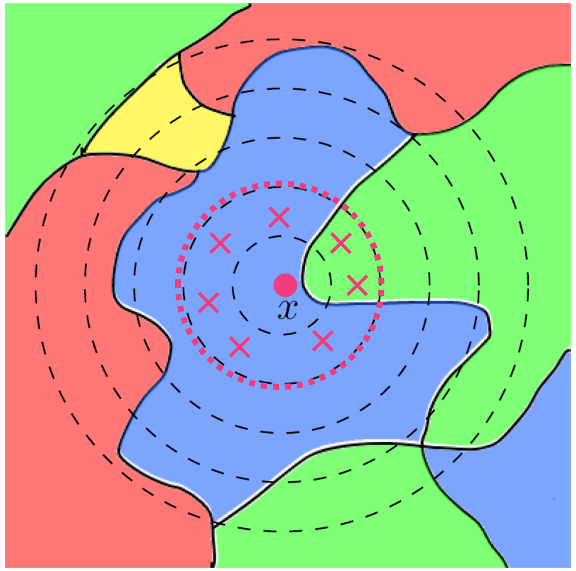
Challenge: DNNs are too complex to regularize every $f(x+\delta)$

- Randomized smoothing (RS) instead constructs another classifier $\hat{f}$ from f

$$\hat{f}(x) := \arg\max_{k\in\mathcal{Y}} \left\{ \mathbb{P}_{\delta\sim\mathcal{N}(0,\sigma^2 I)} \left(f(x+\delta) = k \right) \right\}$$

Gaussian noise





[Cohen et al., 2019] Certified adversarial robustness via randomized smoothing. ICML 2019.

Background: Randomized Smoothing

Challenge: DNNs are too complex to regularize every $f(x+\delta)$

- Randomized smoothing (RS) instead constructs another classifier $\hat{f}$ from f

$$\hat{f}(x) := \arg\max_{k\in\mathcal{Y}} \left\{\mathbb{P}_{\delta\sim\mathcal{N}(0,\sigma^2 I)}\left(f(x+\delta)=k\right)\right\}$$

Gaussian noise

  - Then, $\hat{f}$ is much easier to obtain adversarial robustness
  - **Cohen et al. (2019)**: A provable guarantee on the robust radius of $\hat{f}$ in terms of f

Theorem Let $p_x := \max_k \mathbb{P}_\delta(f(x+\delta)=k)$. Then, the ℓ_2 robust radius of $\hat{f}(x)$ is lower-bounded by:

$$R(\hat{f};x) := \min_{\hat{f}(x+\delta)\neq\hat{f}(x)} \|\delta\|_2 \geq \sigma\cdot\Phi^{-1}(p_x)$$

Gaussian CDF

[Cohen et al., 2019] Certified adversarial robustness via randomized smoothing. ICML 2019.

Background: Randomized Smoothing

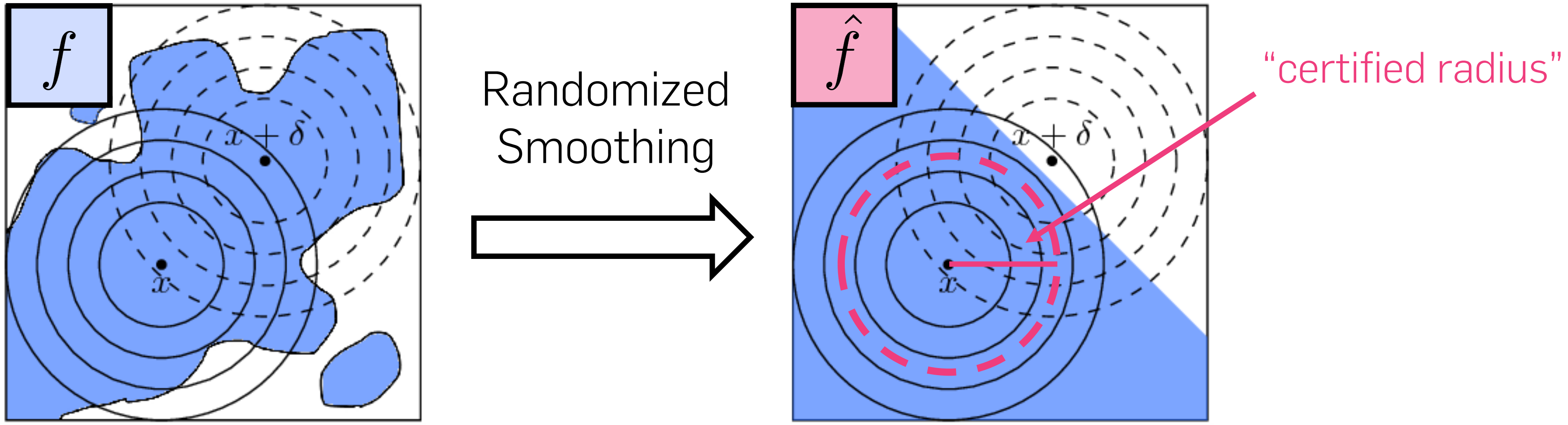
Challenge: DNNs are too complex to regularize every $f(x+\delta)$

- Randomized smoothing (RS) instead constructs another classifier $\hat{f}$ from f

$$\hat{f}(x) := \arg\max_{k \in \mathcal{Y}} \left\{ \mathbb{P}_{\delta \sim \mathcal{N}(0, \sigma^2 I)} \left(f(x+\delta) = k \right) \right\}$$

Gaussian noise

  - Then, $\hat{f}$ is much easier to obtain adversarial robustness



[Cohen et al., 2019] Certified adversarial robustness via randomized smoothing. ICML 2019.

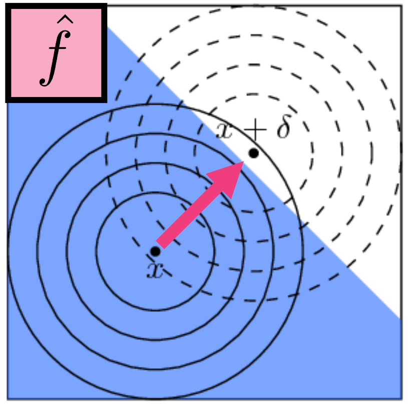
Robust Training for Smoothed Classifiers

🤔 Which f would maximize the robustness of $\hat{f}$?

- **Gaussian** [Cohen et al., 2019]: Training with Gaussian augmentation

$$L^{\mathtt{nat}} := \mathbb{E}_{\delta\sim\mathcal{N}(0,\sigma^2 I)}[\mathcal{L}(F(x+\delta), y)]$$

softmax outputs



More sophisticated training indeed improves certified robustness

- **SmoothAdv** [Salman et al., 2019]: Adversarial training for $\hat{f}$ (approx.)
  - Achieves state-of-the-art certified robustness
- **MACER** [Zhai et al., 2020]: Maximizing a soft approx. of certified radius
- **Consistency** [Jeong and Shin, 2020]: Minimizing the variance of prediction over noise

[Cohen et al., 2019] Certified adversarial robustness via randomized smoothing. ICML 2019.
[Salman et al., 2019] Provably robust deep learning via adversarially trained smoothed classifiers. NeurIPS 2019.
[Zhai et al., 2020] MACER: attack-free and scalable robust training via maximizing certified radius. ICLR 2020.
[Jeong and Shin, 2020] Consistency Regularization for Certified Robustness of Smoothed Classifiers. NeurIPS 2020.

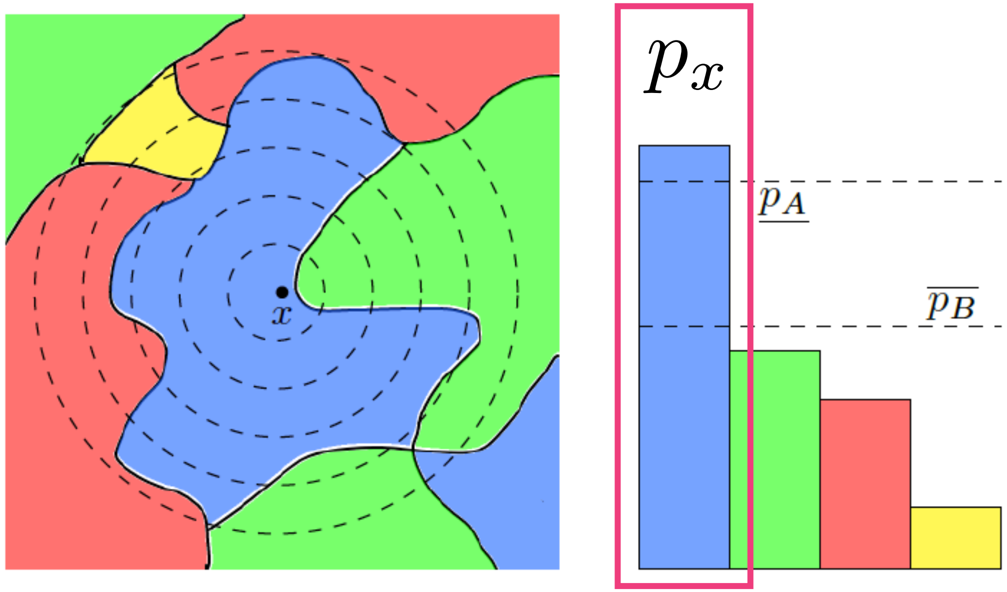
Motivation: Confidence and Robustness in RS

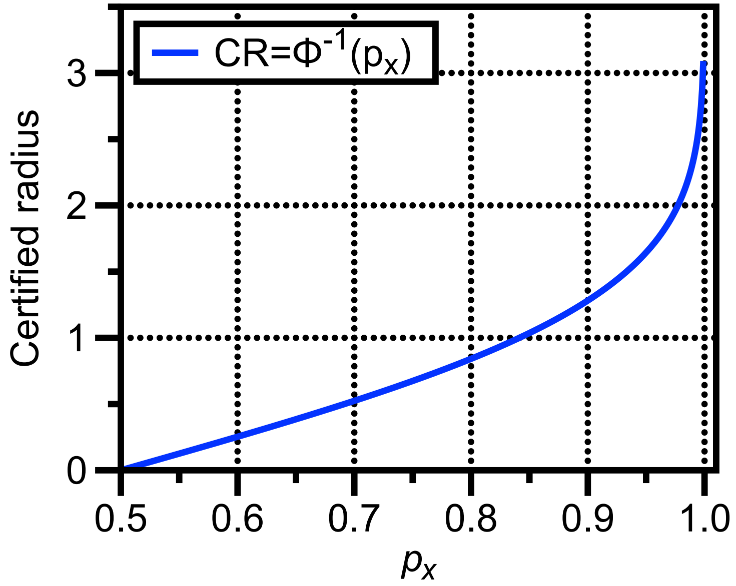
Remark: The prediction confidence p lower-bounds the adversarial robustness of $\hat{f}$

<u>Theorem</u> Let $p_x := \max_k \mathbb{P}_\delta(f(x+\delta)=k)$. Then, the ℓ_2 robust radius of $\hat{f}(x)$ is lower-bounded by:

$$R(\hat{f};x) := \min_{\hat{f}(x+\delta)\neq\hat{f}(x)} \|\delta\|_2 \geq \sigma \cdot \Phi^{-1}(p_x)$$

Gaussian CDF





[Cohen et al., 2019] Certified adversarial robustness via randomized smoothing. ICML 2019.

Motivation: Confidence and Robustness in RS

Remark: The prediction confidence p lower-bounds the adversarial robustness of $\hat{f}$

<u>Theorem</u> Let $p_x := \max_k \mathbb{P}_\delta(f(x+\delta)=k)$. Then, the ℓ_2 robust radius of $\hat{f}(x)$ is lower-bounded by:

$$R(\hat{f};x) := \min_{\hat{f}(x+\delta)\neq\hat{f}(x)} \|\delta\|_2 \geq \sigma \cdot \Phi^{-1}(p_x)$$

Gaussian CDF

- The higher p_x, the better robustness at x
- Standard (non-smoothed) DNNs do not have this property

🤔 Will a better confidence calibration bring a more robust $\hat{f}$?

- Do current smoothed classifiers "well-calibrated" for unseen inputs?
- If not, how will such inputs affect the (certified) robustness of $\hat{f}$?

SmoothMix: Confidence-calibrated Training of RS

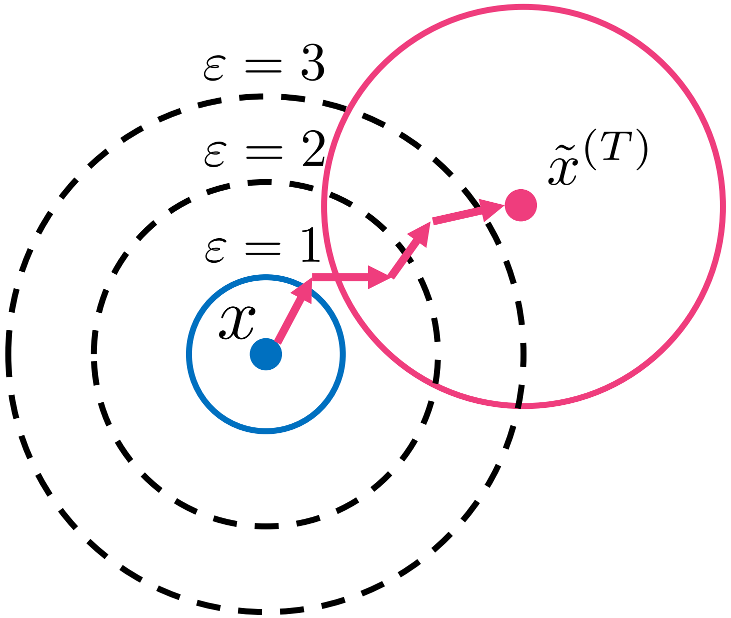
Observation: $\hat{f}$ is often over-confident at nearby, off-class inputs of x

- Such inputs can negatively affect the robustness at x
  - Due to the relationship: **higher confidence → better robustness**

over-confidence

CIFAR-10 (Test set; %)	Clean	$\varepsilon = 1.0$	$\varepsilon = 2.0$	$\varepsilon = 3.0$	$\varepsilon = 4.0$	$\varepsilon = 5.0$
$\mathbb{E}[\mathbb{P}(f(x+\delta) = y)]$	**66.4**	47.1	24.3	14.2	11.3	10.7
$\mathbb{E}[\max_{c \neq y} \mathbb{P}(f(x+\delta) = c)]$	24.2	37.8	59.5	**71.8**	**78.5**	**82.0**

Max. off-class confidence



- An unrestricted adversarial search can effectively find the "over-confident" inputs:

$$\tilde{x}^{(t+1)} := \tilde{x}^{(t)} + \alpha \cdot \frac{\nabla_x J(\tilde{x}^{(t)})}{\|\nabla_x J(\tilde{x}^{(t)})\|_2}, \text{ where } J(x) := -\log\left(\frac{1}{m}\sum_i F_y(x+\delta_i)\right)$$

SmoothMix: Confidence-calibrated Training of RS

Observation: $\hat{f}$ is often over-confident at nearby, off-class inputs of x

How can we effectively control the confidence of $\tilde{x}^{(T)}$ while keeping those of x ?

Mix-Up training [Zhang et al. 2018] between x and $\tilde{x}^{(T)}$

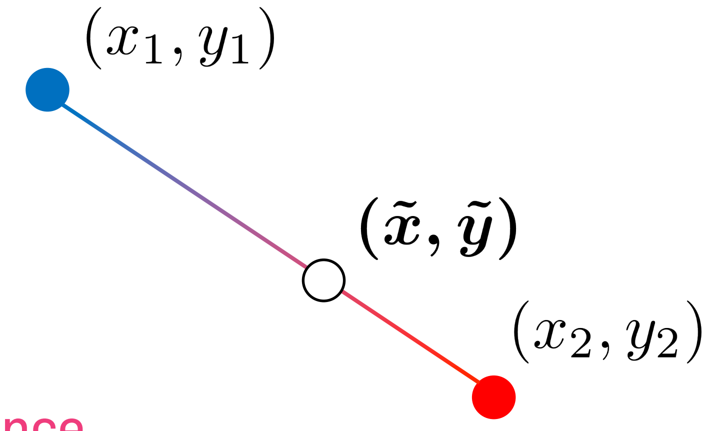
- It keeps the original confidence at x of $\hat{F}_y(x) = \frac{1}{m}\sum_{i=1}^{m} F_y(x)$
- The over-confident input $\tilde{x}^{(T)}$ is regularized toward the uniform confidence

$$L^{\mathtt{mix}} := \mathbb{E}_{\delta\sim\mathcal{N}(0,\sigma^2 I)}\left[\mathcal{L}(F(x^{\mathtt{mix}}+\delta), y^{\mathtt{mix}})\right]$$

$$x^{\mathtt{mix}} := (1-\lambda)\cdot x + \lambda\cdot\tilde{x}^{(T)}$$

$$y^{\mathtt{mix}} := (1-\lambda)\cdot\hat{F}(x) + \lambda\cdot\frac{1}{C}$$

"uniform" confidence



[Zhang et al., 2017] mixup: Beyond Empirical Risk Minimization, ICLR 2018.

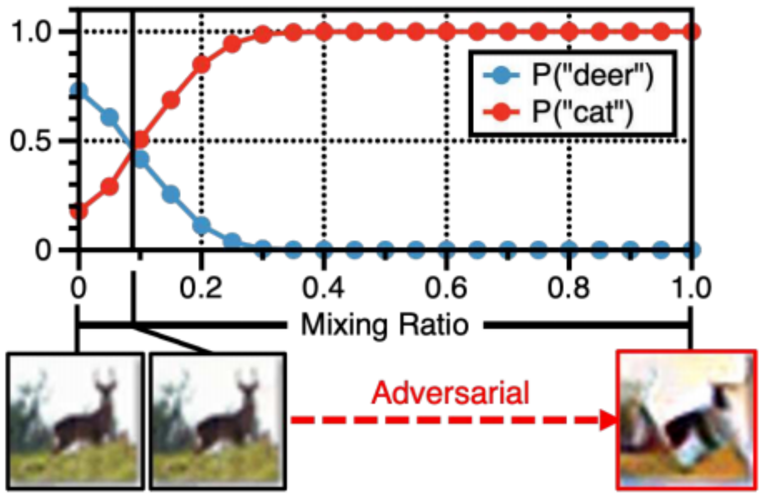
SmoothMix: Confidence-calibrated Training of RS

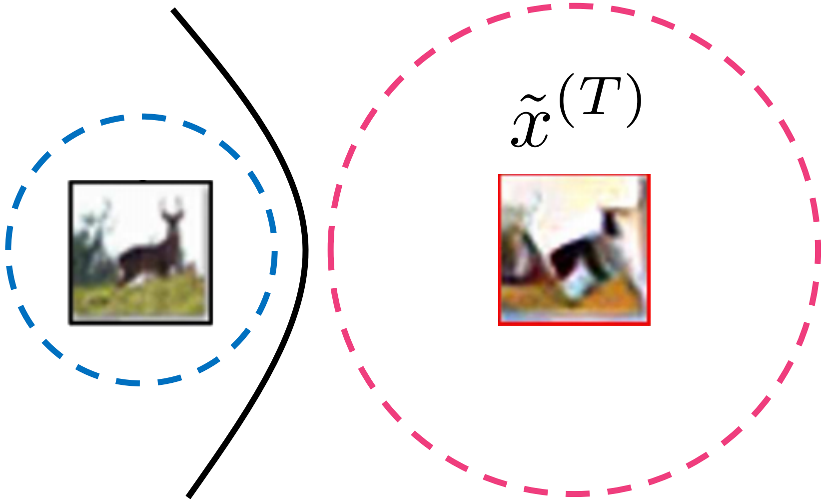
Observation: $\hat{f}$ is often over-confident at nearby, off-class inputs of x

🤔 How can we effectively control the confidence of $\tilde{x}^{(T)}$ while keeping those of x ?

💡 Mix-Up training [Zhang et al. 2018] between x and $\tilde{x}^{(T)}$

- It keeps the original confidence at x of $\hat{F}_y(x) = \frac{1}{m}\sum_{i=1}^{m} F_y(x)$
- The over-confident input $\tilde{x}^{(T)}$ is regularized toward the uniform confidence





[Zhang et al., 2017] mixup: Beyond Empirical Risk Minimization, ICLR 2018.

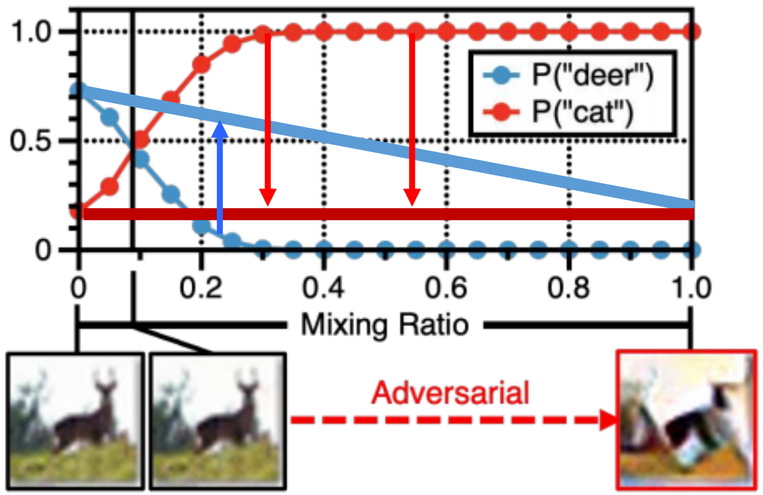
SmoothMix: Confidence-calibrated Training of RS

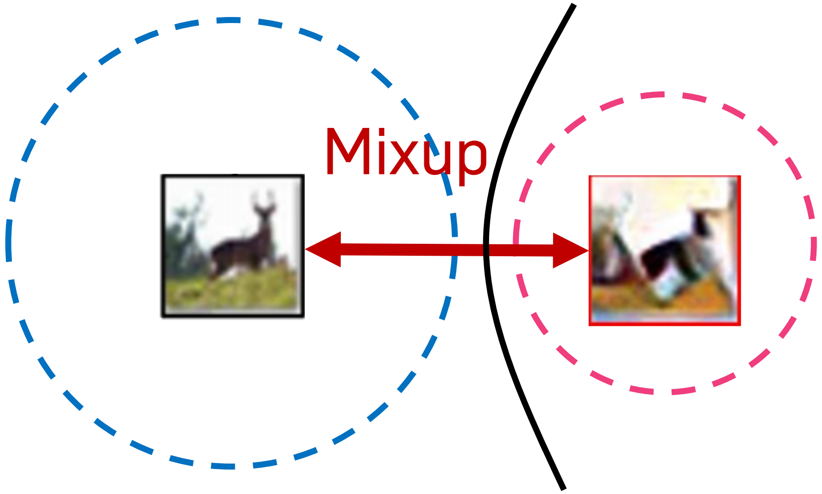
Observation: $\hat{f}$ is often over-confident at nearby, off-class inputs of x

🤔 How can we effectively control the confidence of $\tilde{x}^{(T)}$ while keeping those of x ?

💡 Mix-Up training [Zhang et al. 2018] between x and $\tilde{x}^{(T)}$

- It keeps the original confidence at x of $\hat{F}_y(x) = \frac{1}{m}\sum_{i=1}^{m} F_y(x)$
- The over-confident input $\tilde{x}^{(T)}$ is regularized toward the uniform confidence





[Zhang et al., 2017] mixup: Beyond Empirical Risk Minimization, ICLR 2018.

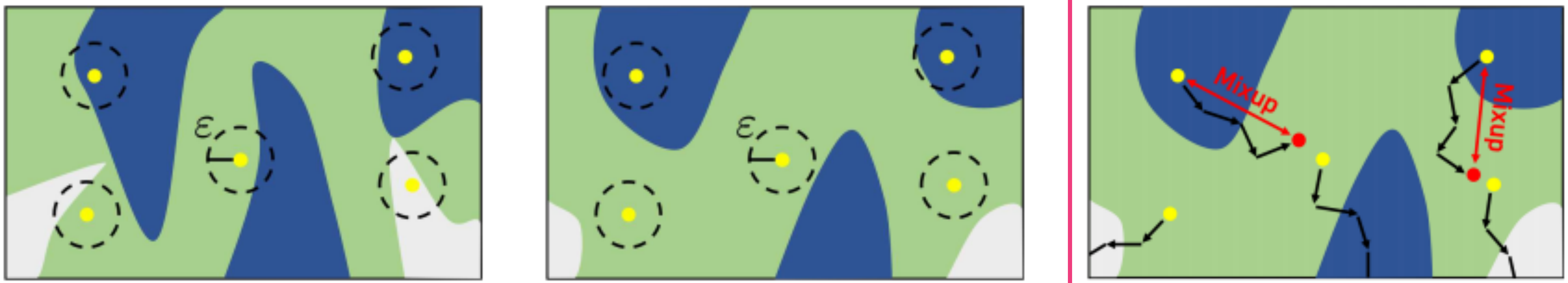
SmoothMix: Confidence-calibrated Training of RS

SmoothAdv [Salman et al., 2019]: Applying AT for $\hat{f}$ can improve RS

- AT assumes a hard ε-ball – RS may already offer the robustness under this constraint

$$\hat{x} = \arg\max_{\|x'-x\|_2 \le \epsilon} \mathcal{L}(\hat{F}; x', y) \approx \arg\max_{\|x'-x\|_2 \le \epsilon} \left(-\log \left(\frac{1}{m} \sum_i F_y(x' + \delta_i) \right) \right)$$

SmoothMix proposes an "unrestricted" way to apply AT for smoothed classifiers



(a) Adversarial training [40] (b) SmoothAdv [49] (c) SmoothMix (Ours)

[Salman et al., 2019] Provably robust deep learning via adversarially trained smoothed classifiers. NeurIPS 2019.

SmoothMix: Confidence-calibrated Training of RS

SmoothMix = A new AT method specially designed for RS

1. Unrestrictive search of adversarial examples (AEs)
   - Focuses on finding nearby off-class, but over-confident inputs
2. Minimizes the mixup loss between Clean & AE
   - The AEs are regularized toward the uniform confidence

The final loss of SmoothMix is given by:

$$L := L^{\texttt{nat}} + \eta \cdot L^{\texttt{mix}}$$

- Natural loss: $L^{\texttt{nat}} := \mathbb{E}_{\delta}\left[\mathcal{L}(F(x+\delta), y)\right]$
- Robust loss: $L^{\texttt{mix}} := \mathbb{E}_{\delta \sim \mathcal{N}(0,\sigma^2 I)}\left[\mathcal{L}(F(x^{\texttt{mix}}+\delta), y^{\texttt{mix}})\right]$
- $\eta > 0$: a hyperparameter to control the trade-off between accuracy & robustness

Experimental Results

We evaluate ℓ_2 certified robustness of various training methods for RS:

- Gaussian augmentation [Cohen et al., 2019]
- SmoothAdv [Salman et al., 2019]
- Stability training [Li et al., 2019]
- MACER [Zhai et al., 2020]
- Consistency [Jeong and Shin, 2020]

Evaluation metrics

1. **Certified test accuracy @ radius r** [Cohen et al., 2019]
   - % test dataset that (a) $\hat{f}(x) = y$, and (b) $\mathrm{CR}(f, \sigma, x) := \sigma \cdot \Phi^{-1}(p_A) > r$
2. **Average certified radius (ACR)** [Zhai et al., 2020]

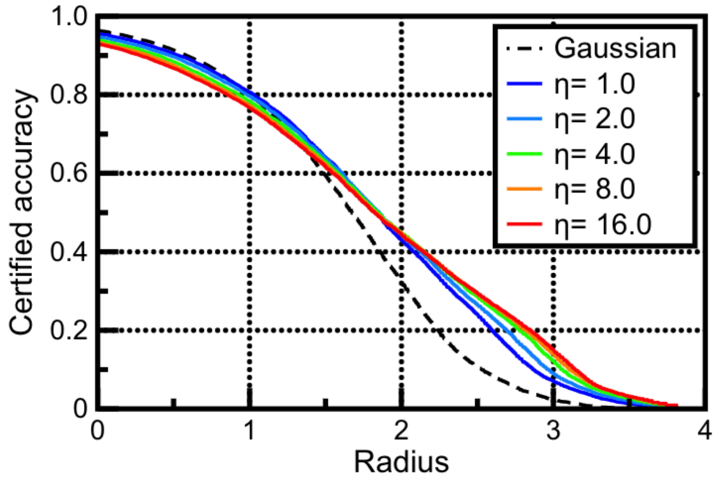
$$\mathrm{ACR} := \frac{1}{|\mathcal{D}_{\texttt{test}}|} \sum_{(x,y)\in\mathcal{D}_{\texttt{test}}} \mathrm{CR}(f, \sigma, x) \cdot \mathbf{1}_{\hat{f}(x)=y}$$

[Cohen et al., 2019] Certified adversarial robustness via randomized smoothing. ICML 2019.
[Salman et al., 2019] Provably robust deep learning via adversarially trained smoothed classifiers. NeurIPS 2019.
[Li et al., 2019] Certified adversarial robustness with additive noise. NeurIPS 2019.
[Zhai et al., 2020] MACER: attack-free and scalable robust training via maximizing certified radius. ICLR 2020.
[Jeong and Shin, 2020] Consistency Regularization for Certified Robustness of Smoothed Classifiers. NeurIPS 2020.

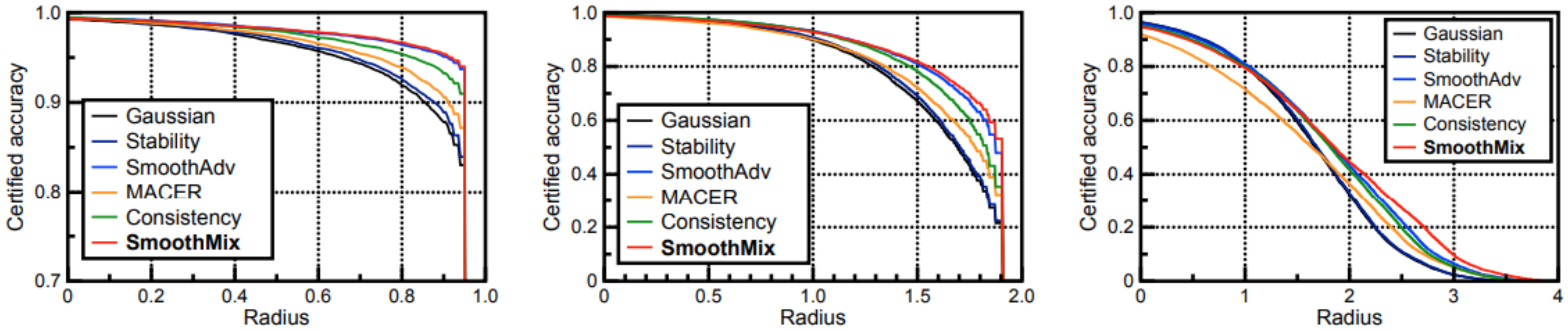
Experimental Results

Results on MNIST

- SmoothMix consistently improves ACR
- The trends hold for a wide range of $\sigma \in \{0.25, 0.5, 1.0\}$
- Shows better trade-offs compared to, e.g., SmoothAdv
- η effectively controls the trade-off: Accuracy ↔ Robustness



Certified accuracy @ varying η



(a) $\sigma = 0.25$ (b) $\sigma = 0.50$ (c) $\sigma = 1.00$

Certified accuracy @ radius $\boldsymbol{r}$

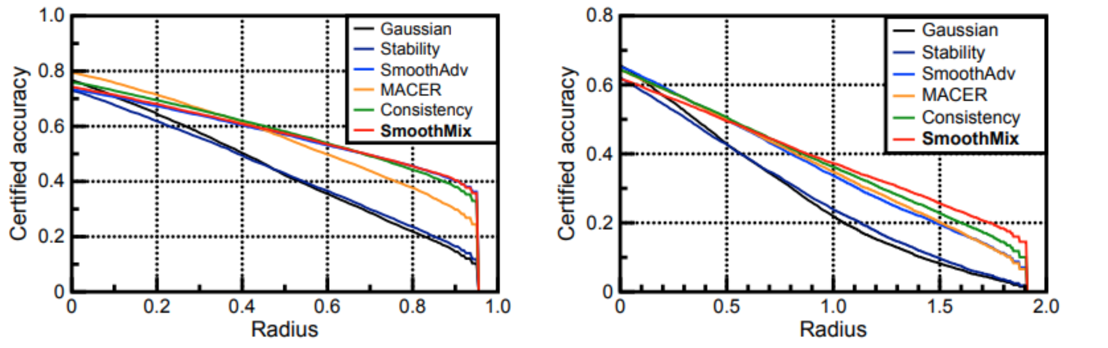
Experimental Results

Results on CIFAR-10 / ImageNet

- The proposed method successfully scales up to ImageNet dataset
- Still exhibits better trade-offs between accuracy and certified robustness

σ	Models (ImageNet)	ACR	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5
0.50	Gaussian (Cohen et al., 2019)	0.733	57	46	37	29	0	0	0	0
	Consistency (Jeong & Shin, 2020)	0.822	55	50	44	34	0	0	0	0
	SmoothAdv (Salman et al., 2019)	0.825	54	49	43	37	0	0	0	0
	SmoothMix (Ours)	**0.846**	55	**50**	**43**	**38**	0	0	0	0
1.00	Gaussian (Cohen et al., 2019)	0.875	44	38	33	26	19	15	12	9
	Consistency (Jeong & Shin, 2020)	0.982	41	37	32	28	24	21	17	14
	SmoothAdv (Salman et al., 2019)	1.040	40	37	34	30	27	25	20	15
	SmoothMix (Ours)	**1.047**	40	37	**34**	**30**	**26**	**24**	**20**	**17**

σ	Models (CIFAR-10)	ACR	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75
0.25	Gaussian (Cohen et al., 2019)	0.424	76.6	61.2	42.2	25.1	0.0	0.0	0.0	0.0
	Stability training (Li et al., 2019)	0.421	72.3	58.0	43.3	27.3	0.0	0.0	0.0	0.0
	SmoothAdv* (Salman et al., 2019)	0.544	73.4	65.6	57.0	47.5	0.0	0.0	0.0	0.0
	MACER* (Zhai et al., 2020)	0.531	79.5	69.0	55.8	40.6	0.0	0.0	0.0	0.0
	Consistency (Jeong & Shin, 2020)	0.552	75.8	67.6	58.1	46.7	0.0	0.0	0.0	0.0
	SmoothMix (Ours)	**0.553**	**77.1**	**67.9**	**57.9**	**46.7**	0.0	0.0	0.0	0.0
	+ One-step adversary	**0.548**	74.2	**66.1**	**57.4**	**47.7**	0.0	0.0	0.0	0.0
0.50	Gaussian (Cohen et al., 2019)	0.525	65.7	54.9	42.8	32.5	22.0	14.1	8.3	3.9
	Stability training (Li et al., 2019)	0.521	60.6	51.5	41.4	32.5	23.9	15.3	9.6	5.0
	SmoothAdv* (Salman et al., 2019)	0.684	65.3	57.8	49.9	41.7	33.7	26.0	19.5	12.9
	MACER* (Zhai et al., 2020)	0.691	64.2	57.5	49.9	42.3	34.8	27.6	20.2	12.6
	Consistency (Jeong & Shin, 2020)	0.720	64.3	57.5	50.6	43.2	36.2	29.5	22.8	16.1
	SmoothMix (Ours)	**0.715**	65.0	**56.7**	**49.2**	**41.2**	**34.5**	**29.6**	**23.5**	**18.1**
	+ One-step adversary	**0.737**	61.8	**55.9**	**49.5**	**43.3**	**37.2**	**31.7**	**25.7**	**19.8**



(a) $\sigma = 0.25$ (b) $\sigma = 0.50$

Certified test accuracy @ radius $\boldsymbol{r}$

Summary

We propose a new form of adversarial training for RS

- It leverages "Confidence → Robustness" in the world of RS
- Nearby, over-confident inputs may harm the robustness of in-distribution samples
- A mixup-based loss could effectively calibrate these over-confident inputs

Randomized smoothing has a great potential toward reliable deep learning

- RS gives a provable guarantee on adversarial robustness
- It also offers an easier & attack-free way to train a robust model than AT
- We hope our work could be a step toward reducing the gap between RS and AT

Please drop by our poster session for more information!