Learning to Generate Noise for Multi-Attack Robustness

Divyam Madaan[1], Jinwoo Shin[2,3] and Sung Ju Hwang[1,3,4]

[1]School of Computing, KAIST, Daejeon, South Korea

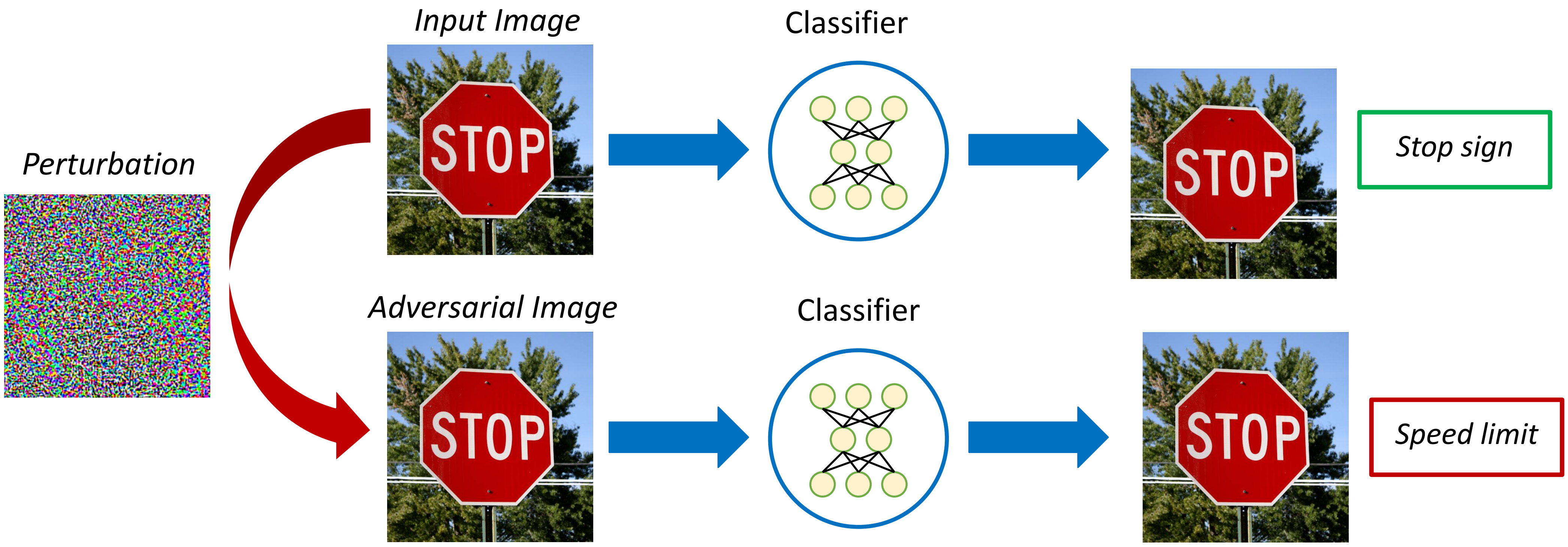
[2]School of Electrical Engineering, KAIST, Daejeon, South Korea

[3]Graduate School of AI, KAIST, Daejeon, South Korea

[4]AITRICS, Seoul, South Korea

Motivation

Adversarial examples are carefully crafted ***imperceptible examples*** for misclassification.



Robustness and accuracy of these networks is important for their deployment in ***safety-critical applications***.

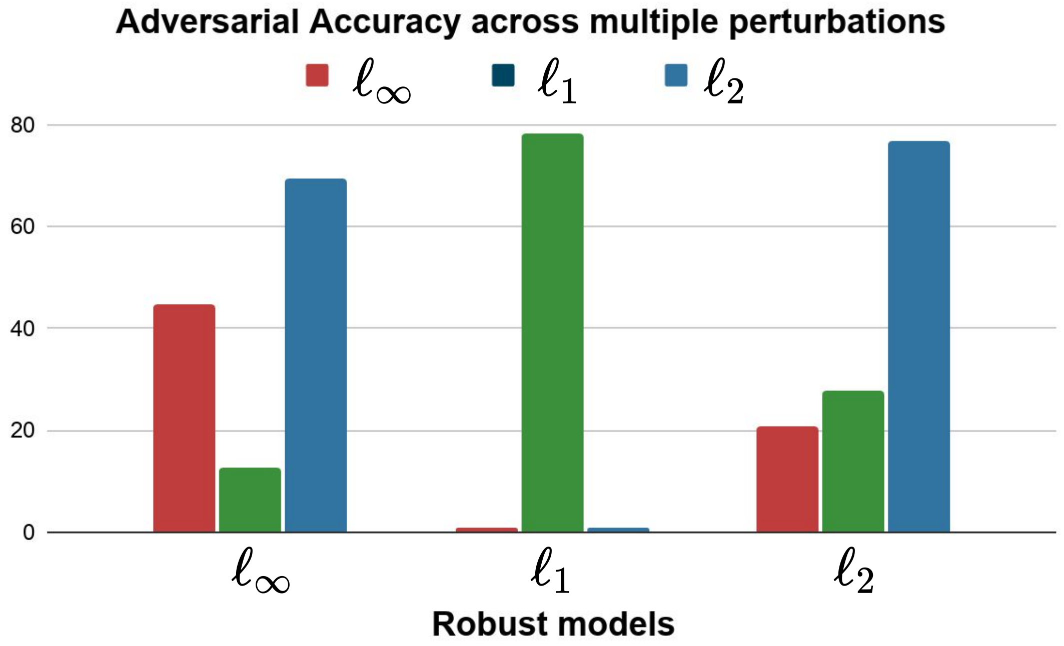
Motivation

The standard adversarial training optimizes the network using a ***min-max formulation*** on a ***single perturbation***.

Algorithm 1 PGD adversarial training

input Dataset $\mathcal{D}$, N epochs, network f_θ with parameters θ, loss function $\mathcal{L}$, attack radius ϵ, step size α for some norm ball $\mathcal{B}$, and T PGD steps
output Final model paramaters θ

```
for n = {1, ..., N} do
    Sample mini-batch (x, y) of size B
    x^adv = x
    // PGD adversarial attack
    for t = {1, ..., T} do
        δ = argmax_{||v|| ≤ α} v^T ∇L(x_adv, y)
        x^adv_{t+1} = proj_{B(x,ε)} (x^adv_t + δ)
    end for
    θ = θ − ∇_θ L(f_θ(x^adv), y)
end for
```



However, single perturbation adversarial training is ***not robust*** against multiple perturbations.

[Madry et al., 2018] Towards Deep Learning Models Resistant to Adversarial Attacks, ICLR 2018

Related Work: Multi Perturbation Adversarial Training

Tramer et al. (2019) proposed optimization with the ***worst/union*** of all the perturbations.

1. Optimize the outer objective with strongest perturbation.

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\operatorname{argmax}_k \mathcal{L}_{\text{cls}} \left(f_\theta \left(\mathcal{A}_k \left(x \right) \right), y \right) \right]$$

2. Optimize the outer objective with all the perturbations.

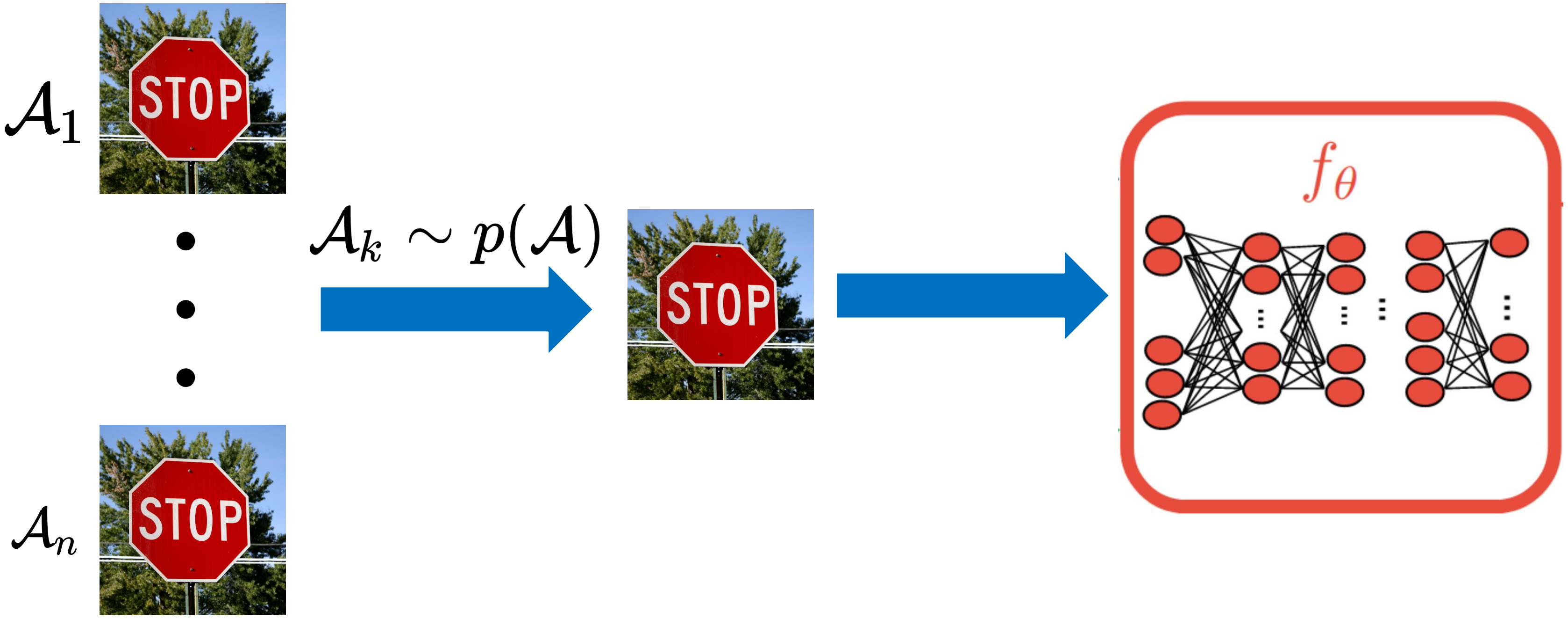
$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \frac{1}{n} \sum_{k=1}^{k=n} \mathcal{L}_{\text{cls}} \left(f_\theta \left(\mathcal{A}_k \left(x \right), y \right) \right)$$

However, multiple perturbation training ***increases*** the training cost by a ***factor of four*** over single perturbation adversarial training.

[Tramer et al., 2019] Adversarial Training and Robustness for Multiple Perturbations, NeurIPS 2019

Stochastic Adversarial Training (SAT)

Our proposed SAT samples from a ***distribution of attacks*** during each episode of training, which ***prevents overfitting*** on a particular perturbation.

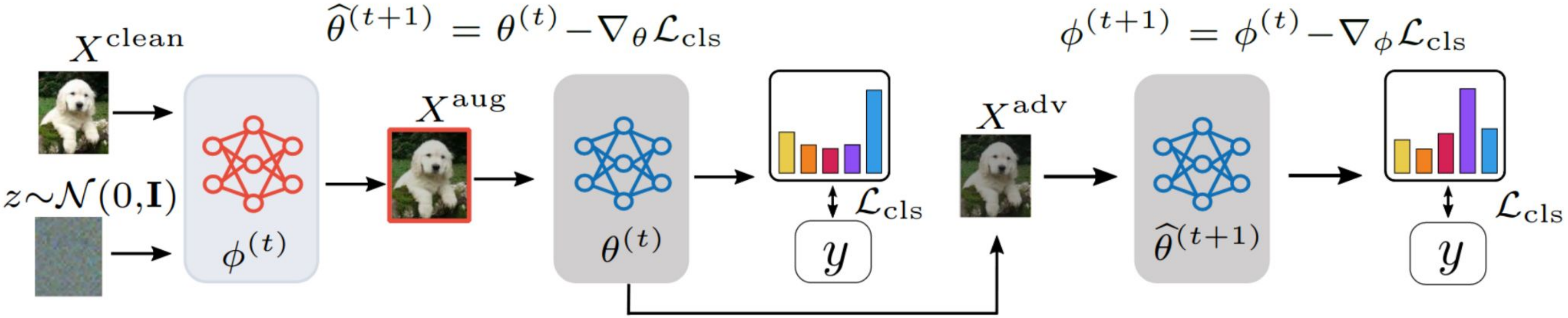


Training objective:

$$\min_{\theta} \mathbb{E}_{\substack{(x,y)\sim\mathcal{D} \\ \mathcal{A}_k \sim p(\mathcal{A})}} \mathcal{L}_{\text{cls}}\left(f_\theta\left(\mathcal{A}_k(x), y\right)\right)$$

Meta-Noise Generator with Adversarial Consistency

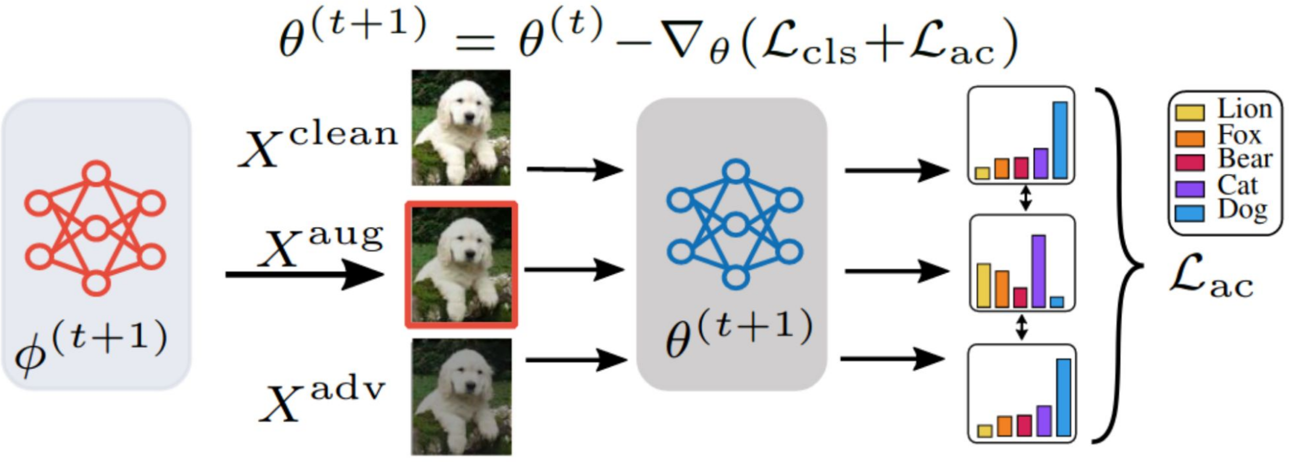
MNG-AC meta-learns to generate ***input-dependent stochastic noise*** to improve model's robustness and ***adversarial consistency*** across multiple attacks.



Step 1: Meta-learn the ***input dependent noise-generator*** $\phi^{(t)}$ to defend against multiple adversarial perturbations.

Meta-Noise Generator with Adversarial Consistency

MNG-AC meta-learns an ***input-dependent stochastic noise distribution*** to improve model's robustness and ***adversarial consistency*** across multiple attacks.



Step 2: Update the classifier $\theta^{(t)}$ with the ***stochastic adversarial loss*** and ***adversarial consistency*** regularization.

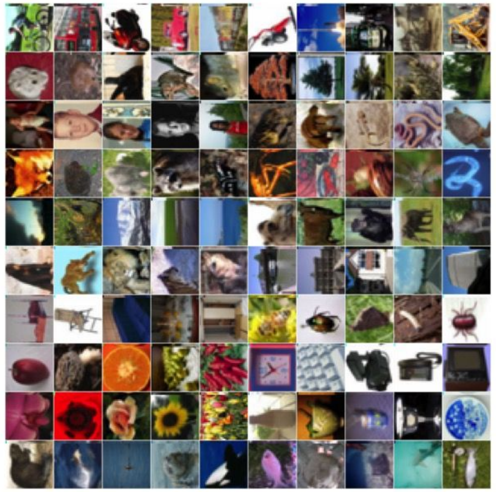
Dataset

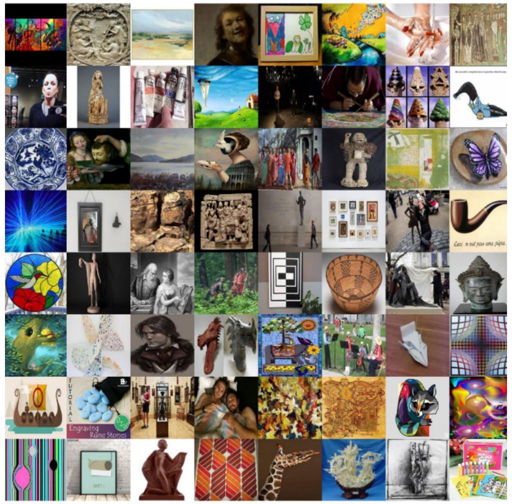
We evaluate our model and baselines on ***three benchmark datasets***.

CIFAR10 [Krizhevsky, 2012] A dataset with 60,000 images from ***ten animal and vehicle classes***.

SVHN [Netzer et al., 2011] A dataset with 99289 of ten digits and numbers classes from ***natural scene images***.

Tiny-ImageNet [Russakovsky, 2015] A ***subset of ImageNet*** dataset with 200 classes.





[Krizhevsky, 2012] Learning multiple layer of features from tiny images. University of Toronto 2012
[Netzer et al, 2012] Reading digits in natural images with unsupervised feature learning. Workshop on Deep Learning and Unsupervised Feature Learning, NeurIPS, 2011
[Ruakovsky, 2015] Imagenet large scale visual recognition challenge. International journal of computer vision, 2015

Result on CIFAR-10 dataset

Our proposed ***MNG-AC*** outperforms the SOTA single-perturbation baselines.

Model	$\text{Acc}_{\text{clean}}$	ℓ_∞	ℓ_1	ℓ_2	$\text{Acc}_{\text{adv}}^{\text{union}}$	$\text{Acc}_{\text{adv}}^{\text{avg}}$	Time (h)
Nat	**94.7**	0.0	4.4	19.4	0.0	7.9	**0.4**
Adv_∞	86.8	44.9	12.8	69.3	12.9	42.6	4.5
Adv_1	**93.3**	0.0	**78.1**	0.0	0.0	25.1	8.1
Adv_2	91.7	20.7	27.7	**76.8**	17.9	47.6	**3.7**
TRADES_∞	84.7	**48.9**	17.9	69.4	17.2	45.4	5.2
MNG-AC	81.5	42.2	**55.0**	71.5	**41.6**	**56.2**	11.2

[Zhang et al., 2019] Theoretically Principled Trade-off between Robustness and Accuracy. ICML 2019

Result on CIFAR-10 dataset

Our proposed ***MNG-AC*** outperforms the SOTA multi-perturbation baselines.

Model	$\mathrm{Acc}_{\mathrm{clean}}$	ℓ_∞	ℓ_1	ℓ_2	$\mathrm{Acc}_{\mathrm{adv}}^{\mathrm{union}}$	$\mathrm{Acc}_{\mathrm{adv}}^{\mathrm{avg}}$	Time (h)
Nat	**94.7**	0.0	0.0	0.4	0.0	0.0	**0.4**
Adv_∞	86.8	**44.9**	26.2	55.0	25.6	41.9	4.5
Adv_1	**93.3**	0.0	**80.7**	0.0	0.0	26.8	8.1
Adv_2	89.4	28.8	54.2	**65.8**	28.6	49.6	**3.7**
TRADES_∞	84.7	**48.9**	32.3	57.8	31.5	46.3	5.2
$\mathrm{Adv}_{\mathrm{avg}}$	86.0	34.1	61.3	**65.7**	34.1	**53.7**	16.9
$\mathrm{Adv}_{\mathrm{max}}$	84.2	39.9	57.9	64.5	39.7	54.1	16.3
MSD	82.7	43.5	54.3	63.1	**42.7**	53.6	16.7
MNG-AC	81.7	41.4	**65.4**	65.2	**41.4**	**57.2**	8.4

[Tramer et al., 2019] Adversarial Training and Robustness for Multiple Perturbations, NeurIPS 2019
[Maini et al., 2020] Adversarial Robustness Against the Union of Multiple Perturbation Models, ICML 2020

KAIST

AITRICS

Result on SVHN dataset

Our proposed ***MNG-AC*** outperforms the SOTA multi-perturbation baselines.

Model	$\text{Acc}_{\text{clean}}$	ℓ_∞	ℓ_1	ℓ_2	$\text{Acc}_{\text{adv}}^{\text{union}}$	$\text{Acc}_{\text{adv}}^{\text{avg}}$	Time (h)
Nat	**96.8**	0.0	9.4	3.8	0.0	4.5	**0.6**
Adv_∞	92.8	**46.2**	8.2	30.2	8.1	28.3	6.2
Adv_1	92.4	0.0	**77.2**	0.0	0.0	25.7	11.8
Adv_2	93.0	21.7	44.7	**62.9**	21.0	43.1	**6.1**
TRADES_∞	**93.9**	**49.9**	4.2	26.7	4.1	26.9	7.9
Adv_{avg}	91.6	21.5	61.2	56.1	20.4	**45.9**	24.1
Adv_{max}	86.9	28.8	48.9	56.3	28.8	44.7	22.7
MSD	81.8	34.1	43.4	54.1	**34.1**	44.0	23.7
MNG-AC	92.6	34.2	**71.3**	**66.7**	**34.2**	**57.4**	11.9

Results with Semi-Supervised Learning

Our proposed MNG-AC ***corroborates semi-supervised learning*** to improve robustness.

Model	$\text{Acc}_{\text{clean}}$	ℓ_∞	ℓ_1	ℓ_2	$\text{Acc}_{\text{adv}}^{\text{union}}$	$\text{Acc}_{\text{adv}}^{\text{avg}}$	Time
			CIFAR-10 dataset				
RST_∞	**88.9**	54.9	36.0	**59.5**	35.7	50.1	**73.5**
MNG-AC	81.7	41.4	65.4	65.2	41.4	57.2	8.4
MNG-AC + RST_∞	88.7	**47.2**	**73.8**	73.7	**47.2**	**64.9**	78.5
			SVHN dataset				
RST_∞	95.6	60.9	3.5	28.8	3.5	31.1	**81.0**
MNG-AC	92.6	34.2	71.3	66.7	34.2	57.4	11.9
MNG-AC + RST_∞	**96.3**	**43.8**	**78.9**	**72.6**	**43.8**	**65.1**	85.0

[Carmon et al., 2019] Unlabeled Data Improves Adversarial Robustness. NeurIPS 2019

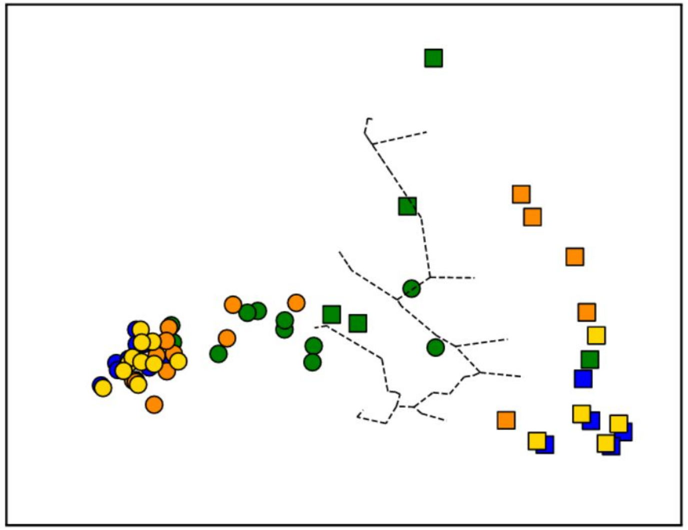
Component analysis

Our proposed MNG-AC ***significantly improves*** the SAT and AC baselines.

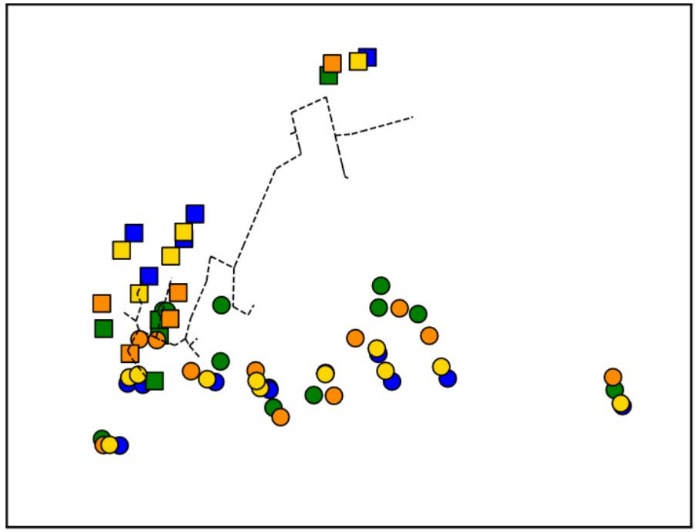
SAT	AC	MNG	$\text{Acc}_{\text{clean}}$	ℓ_∞	ℓ_1	ℓ_2	$\text{Acc}_{\text{adv}}^{\text{union}}$	$\text{Acc}_{\text{adv}}^{\text{avg}}$	Time
					CIFAR-10 dataset				
✔			**86.6**	35.1	61.8	**66.9**	35.0	54.6	**5.5**
✔	✔		80.3	40.6	62.0	63.5	40.6	55.4	6.8
✔	✔	✔	81.7	**41.4**	**65.2**	65.4	**41.4**	**57.2**	8.4
					SVHN dataset				
✔			92.3	26.2	64.4	63.2	26.2	51.0	**7.6**
✔	✔		92.2	31.4	65.2	63.9	31.1	53.5	8.7
✔	✔	✔	**92.6**	**34.2**	**71.3**	**66.7**	**34.2**	**57.4**	11.9

Decision boundary visualization

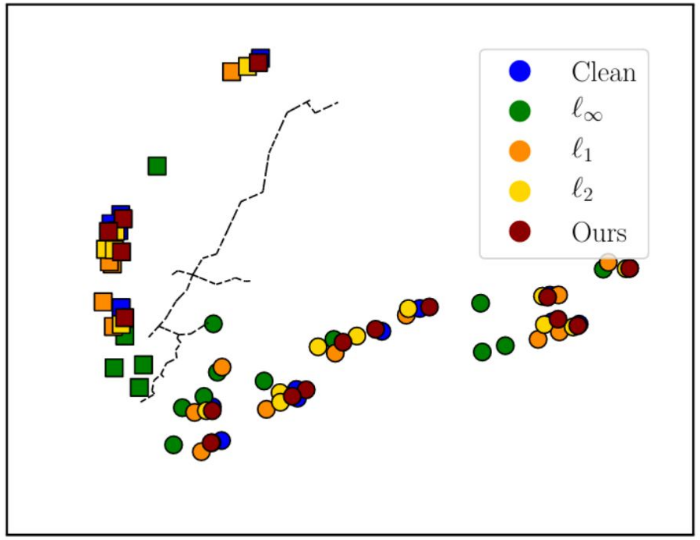
Our proposed MNG-AC ***pushes away the decision boundary*** that in turn improves the overall robustness.



$\mathrm{Adv}_{\mathrm{avg}}$



$\mathrm{Adv}_{\max}$



MNG-AC

Conclusion

- We tackle the problem of ***robustness against multiple perturbations*** and the ***computational overhead*** incurred during multiple perturbations training.
- Meta Noise Generator with Adversarial Consistency (MNG-AC) explicitly meta-learns an ***input-dependent noise*** to minimize the ***stochastic adversarial loss*** and promote ***adversarial label consistency*** across multiple attacks.
- Results show that our model ***pushes away the decision boundary, improves robustness against multiple perturbations*** with ***negligible training cost*.**
- We believe that our paper can be a ***strong guideline when other researchers pursue similar tasks in the future***.

Codes and pretrained models available at https://github.com/divyam3897/MNG_AC

Thank you