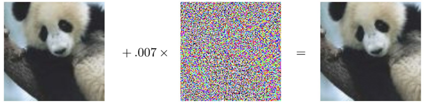
Adversarial Attack Generation Empowered by Min-Max Optimization

Jingkang Wang*, Tianyun Zhang*, Sijia Liu, Pin-Yu Chen, Jiacen Xu, Makan Fardad, Bo Li

Motivation: *can min-max do beyond AT?*

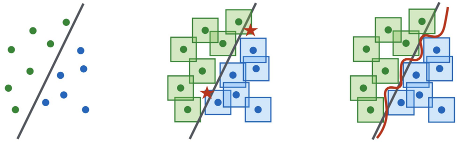
- Neural networks are susceptible to adversarial attacks



Classified as **Panda** | Imperceptible Perturbation | Classified as **Gibbon**

- Adversarial training (AT, Madry et al, 2018):

$$\min_{\theta} \rho(\theta), \quad \text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\delta\in\mathcal{S}} L(\theta, x+\delta, y)\right]$$



- *Beyond AT, can other types of* ***min-max*** *formulation and optimization techniques advance the research in adversarial attack generation?*

Min-Max Across Domains

- Robust optimization over K risk domains (optimize the worst-case performance):

$$\underset{\mathbf{v}\in\mathcal{V}}{\text{minimize}} \ \underset{i\in[K]}{\text{maximize}} \quad F_i(\mathbf{v})$$

$$\Updownarrow$$

$$\underset{\mathbf{v}\in\mathcal{V}}{\text{minimize}} \ \underset{\mathbf{w}\in\mathcal{P}}{\text{maximize}} \quad \sum_{i=1}^{K} w_i F_i(\mathbf{v})$$

$$\mathcal{P} = \{\mathbf{w} \,|\, \mathbf{1}^T\mathbf{w} = 1, w_i \in [0,1], \forall i\}$$

non-stable learning!

- One hot representation **reduces the generalizability** to other domains and **induces instability** of the learning procedure

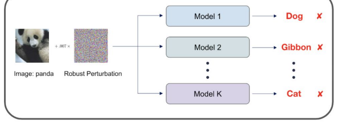
- Regularized Formulation (strike a balance between the average and the worst-case performance):

$$\underset{\mathbf{v}\in\mathcal{V}}{\text{minimize}} \ \underset{\mathbf{w}\in\mathcal{P}}{\text{maximize}} \quad \sum_{i=1}^{K} \underbrace{w_i}_{\text{Domain weights}} F_i(\mathbf{v}) \underbrace{- \tfrac{\gamma}{2}\|\mathbf{w} - \mathbf{1}/K\|_2^2}_{\text{strongly concave regularizer}}$$

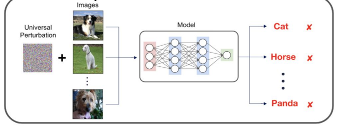
Min-Max Power in Attack Design

- We can design the unified min-max framework actually fits into **various attack** settings!

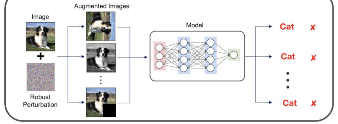
Attacking model ensembles



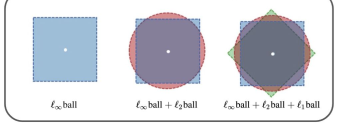
Universal perturbations



Robust attack over multiple transformations



Adversarial training under multiple perturbations



Setting 1: Ensemble attack over multiple models

- Consider K ML/DL models $\{\mathcal{M}_i\}_{i=1}^K$, the goal is to find robust adversarial examples that can fool all K models simultaneously

$$\underset{\boldsymbol{\delta}\in\mathcal{X}}{\text{minimize}} \ \underset{\mathbf{w}\in\mathcal{P}}{\text{maximize}} \quad \sum_{i=1}^{K} w_i f(\boldsymbol{\delta}; \mathbf{x}_0, y_0, \mathcal{M}_i) - \tfrac{\gamma}{2}\|\mathbf{w} - \mathbf{1}/K\|_2^2$$

- $\mathbf{w}$ encodes the difficulty level of attacking each model

Setting 2: Universal perturbation over multiple examples

- Consider K natural examples $\{(\mathbf{x}_i, y_i)\}_{i=1}^K$ and a single model $\mathcal{M}$ the goal is to find the universal perturbation δ so that all the corrupted K examples can fool

$$\underset{\boldsymbol{\delta}\in\mathcal{X}}{\text{minimize}} \ \underset{\mathbf{w}\in\mathcal{P}}{\text{maximize}} \quad \sum_{i=1}^{K} w_i f(\boldsymbol{\delta}; \mathbf{x}_i, y_i, \mathcal{M}) - \tfrac{\gamma}{2}\|\mathbf{w} - \mathbf{1}/K\|_2^2$$

- $\mathbf{w}$ encodes the difficulty level of attacking each image

Setting 3: Robust attack over data transformations

- Consider K categories of data transformation $\{p_i\}$e.g., rotation, lightening, and translation. The goal to find the adversarial attack that is robust to data transformations

$$\underset{\boldsymbol{\delta}\in\mathcal{X}}{\text{minimize}} \ \underset{\mathbf{w}\in\mathcal{P}}{\text{maximize}} \quad \sum_{i=1}^{K} w_i \mathbb{E}_{t\sim p_i}[f(t(\mathbf{x}_0 + \boldsymbol{\delta}); y_0, \mathcal{M})] - \tfrac{\gamma}{2}\|\mathbf{w} - \mathbf{1}/K\|_2^2$$

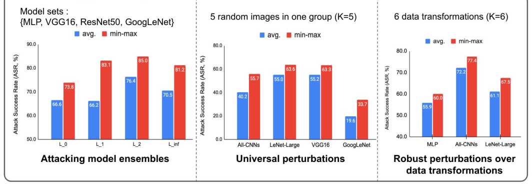
- $\mathbf{w}$ encodes the difficulty level of attacking each type of transformed examples

Results

We produce more robust adversarial attacks

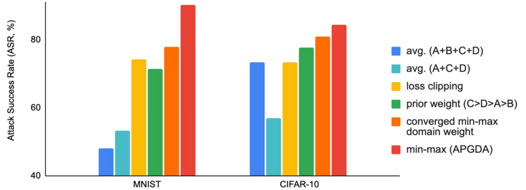
- Significant improvements over average strategy on three robust adversarial attacks

CIFAR-10 (L_inf Attack)



Attacking model ensembles | Universal perturbations | Robust perturbations over data transformations

- Outperforms heuristic strategies in an affordable way!



- A holistic tool to interpret the risk of different domain sources

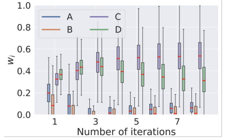
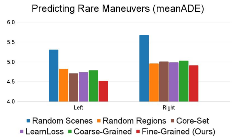


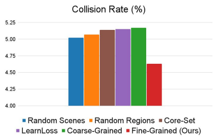
Image						
Weight	ℓ_0	0.	0.	0.	0.	1.000
	ℓ_1	0.	0.	0.	0.	1.000
	ℓ_2	0.	0.	0.	0.	1.000
	ℓ_∞	0.	0.	0.	0.	1.000
Metric	dist.(C&W ℓ_2)	1.839	1.954	1.347	1.698	**3.041**
	$\epsilon_{\min}$ (ℓ_∞)	0.113	0.167	0.073	0.121	**0.199**

Conclusion

- We revisit the strength of min-max optimization in the context of adversarial attack generation.
- Beyond AT, we show that many attack generation or defense problems can be re-formulated in our unified min-max framework
- Our approach results in superior performance and interpretability
- Code is publicly available: github.com/wangjksjtu/minmax-adv

SCAN ME





Selection	Prediction (meanADE) ↓				Downstream Planning				
	Straight (m)	Left (m)	Right (m)	Stationary (m)	Collision ↓ (%)	L2 ↓ (m)	Lat. acc. ↓ (m / s^2)	Jerk ↓ (m / s^3)	Progress ↑ (m)
Random Scenes	2.89	5.31	5.68	0.22	5.02	5.89	2.80	2.67	33.5
Random Regions	2.46	4.82	4.96	**0.20**	5.07	5.71	2.70	2.47	33.6
Core-Set	2.45	4.71	5.01	0.21	5.14	5.72	2.65	2.45	33.6
LearnLoss	2.46	4.74	4.99	0.21	5.15	5.74	2.68	2.47	33.6
Coarse-Grained	2.44	4.79	5.03	0.22	5.17	5.71	2.67	2.44	**33.8**
Fine-Grained	**2.29**	**4.52**	**4.91**	0.21	**4.63**	**5.56**	**2.62**	**2.38**	33.7

1 UNIVERSITY OF TORONTO

2 VECTOR INSTITUTE

3 CLEVELAND STATE UNIVERSITY 1964

4 MICHIGAN STATE UNIVERSITY

5 MIT-IBM Watson AI Lab

6 THE UNIVERSITY OF CALIFORNIA • IRVINE •

7 SYRACUSE ORANGE

8