MITSUBISHI ELECTRIC RESEARCH LABORATORIES (MERL)
Cambridge, Massachusetts

Robust Machine Learning via Privacy/Rate-Distortion Theory

Ye Wang[1], Shuchin Aeron[2], Adnan Siraj Rakin[3], Toshiaki Koike-Akino[1], Pierre Moulin[4]

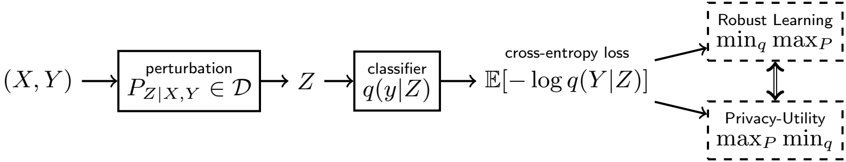
[1]**MERL,** [2]**Tufts University,** [3]**Arizona State University,** [4]**University of Illinois at Urbana-Champaign**

IEEE International Symposium on Information Theory (ISIT 2021)

MITSUBISHI ELECTRIC
Changes for the Better

Connecting Robust ML to Privacy/Rate-Distortion Theory

Motivation: Adversarial Examples, small input perturbations fool deep neural networks



Optimal Privacy-Utility Tradeoff for Data Release [Calmon, Fawaz, 2012]

- Perturbation is *Data Release Mechanism*, Classifier is *Privacy Adversary*
- Mechanism design: maximin problem reduces to max entropy

Robust Machine Learning [Madry et al, 2018]

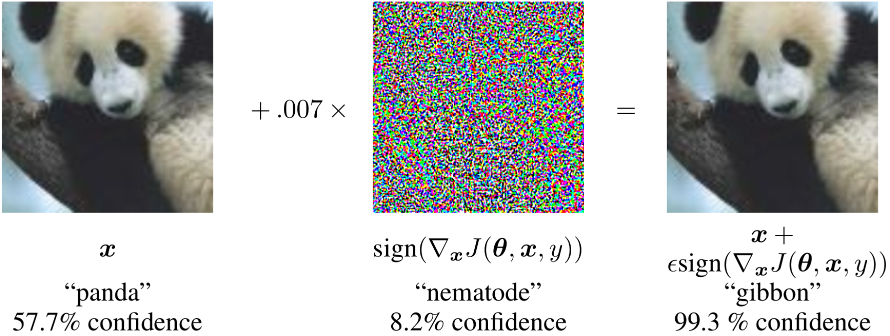
- Classifier is *Robust Model*, Perturbation is *Adversarial Input Attacker*
- Robust model design: minimax solution can be found via max entropy

Similar minimax result of [Tse, Farnia, 2016] limited by technical conditions

Adversarial Examples

Discovered by [Szegedy et al, 2013] in “Intriguing properties of neural networks”

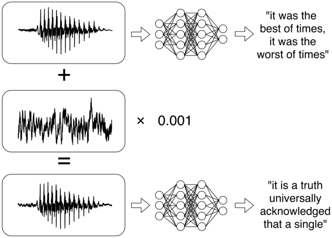
- “Explaining and Harnessing Adversarial Examples” [Goodfellow et al, 2014]



- Small, imperceptible perturbations can fool deep neural networks

Many Other Adversarial Examples

Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%



[Sharif et al, 2016], [Athalye et al, 2018], [Eykholt et al, 2018], [Carlini, Wagner, 2018]

Adversarial Examples Vulnerability in Tesla Auto-Pilot

Tencent Keen Security Lab: first demo of attack on commercial vision product

boingboing / CORY DOCTOROW / 4:06 PM SUN MAR 31, 2019

Small stickers on the ground trick Tesla autopilot into steering into opposing traffic lane

Fig 35. In-car perspective when testing, the red circle marks, the interference markings are marked

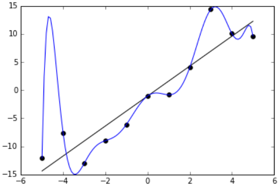
Why do Adversarial Examples Matter?

Besides safety, security, reliability . . .

- Better understanding might yield fundamental insights on machine learning

Potential to broadly impact how we understand and apply ML

- How do we fix broken systems? More data/training? Model depth/architecture?
- What does adversarial fragility imply about generalizability?
- How do we avoid overfitting with highly overparameterized models?



Adversarial examples and defenses are a cat-and-mouse game in the literature

- Fundamental guarantees to break this cycle?

MITSUBISHI ELECTRIC
Changes for the Better

Robust Machine Learning Formulation

Conventional supervised learning formulation: $\min_\theta \mathbb{E}[\ell(f_\theta(X), Y)]$

- Example: classifier $f_\theta(X)$ estimates posterior $q_\theta(y|X)$ over finite label set $\mathcal{Y}$
- Cross-entropy loss: $\ell(f_\theta(X), Y) = -\log q_\theta(Y|X)$
- Note that $\mathbb{E}[-\log q_\theta(Y|X)] = \mathrm{KL}(p_{Y|X}(y|X) \| q_\theta(y|X) | P_X) + H(Y|X)$

MITSUBISHI ELECTRIC
Changes for the Better

Robust Machine Learning Formulation

Conventional supervised learning formulation: $\min_\theta \mathbb{E}[\ell(f_\theta(X), Y)]$

- Example: classifier $f_\theta(X)$ estimates posterior $q_\theta(y|X)$ over finite label set $\mathcal{Y}$
- Cross-entropy loss: $\ell(f_\theta(X), Y) = -\log q_\theta(Y|X)$
- Note that $\mathbb{E}[-\log q_\theta(Y|X)] = \mathrm{KL}(p_{Y|X}(y|X) \| q_\theta(y|X) | P_X) + H(Y|X)$

Robust learning formulation [Madry et al, 2018]

$$\min_\theta \mathbb{E}\left[\max_{\substack{Z \in \mathcal{X}: \\ d(X,Z) \leq \epsilon}} \ell(f_\theta(Z), Y) \right]$$

- Allow perturbations within distance $\epsilon \geq 0$ for some metric $d : \mathcal{X} \times \mathcal{X} \to [0, \infty]$

Generalizing the Robust ML Formulation

$$\min_{\theta} \mathbb{E}\left[\max_{\substack{Z \in \mathcal{X}: \\ d(X,Z) \leq \epsilon}} \ell(f_\theta(Z), Y) \right]$$

MITSUBISHI ELECTRIC
Changes for the Better

Generalizing the Robust ML Formulation

$$\min_{\theta} \mathbb{E}\left[\max_{\substack{Z \in \mathcal{X}: \\ d(X,Z) \leq \epsilon}} \ell(f_\theta(Z), Y) \right]$$

can be reformulated to allow mixed (randomized) strategies for the attacker

$$\min_{\theta} \max_{P_{Z|X,Y} \in \mathcal{D}^*_{d,\epsilon}} \mathbb{E}[\ell(f_\theta(Z), Y)]$$

where the constraint represents the allowable perturbation

$$\mathcal{D}^*_{d,\epsilon} := \{p_{Z|X,Y} \in \mathcal{P}(\mathcal{Z}|\mathcal{X}, \mathcal{Y}) : \Pr[d(X,Z) \leq \epsilon] = 1\}$$

Generalizing the Robust ML Formulation

$$\min_{\theta} \mathbb{E}\left[\max_{\substack{Z \in \mathcal{X}: \\ d(X,Z) \leq \epsilon}} \ell(f_\theta(Z), Y) \right]$$

can be reformulated to allow mixed (randomized) strategies for the attacker

$$\min_{\theta} \max_{P_{Z|X,Y} \in \mathcal{D}^*_{d,\epsilon}} \mathbb{E}[\ell(f_\theta(Z), Y)]$$

where the constraint represents the allowable perturbation

$$\mathcal{D}^*_{d,\epsilon} := \{p_{Z|X,Y} \in \mathcal{P}(\mathcal{Z}|\mathcal{X}, \mathcal{Y}) : \Pr[d(X,Z) \leq \epsilon] = 1\}$$

Alternatively, can strengthen adversary by constraining only expected distortion

$$\mathcal{D}_{d,\epsilon} := \{p_{Z|X,Y} \in \mathcal{P}(\mathcal{Z}|\mathcal{X}, \mathcal{Y}) : \mathbb{E}[d(X,Z)] \leq \epsilon\}$$

Generalizing the Robust ML Formulation

$$\min_{\theta} \mathbb{E}\left[\max_{\substack{Z \in \mathcal{X}: \\ d(X,Z) \le \epsilon}} \ell(f_\theta(Z), Y)\right]$$

can be reformulated to allow mixed (randomized) strategies for the attacker

$$\min_{\theta} \max_{P_{Z|X,Y} \in \mathcal{D}^*_{d,\epsilon}} \mathbb{E}[\ell(f_\theta(Z), Y)]$$

where the constraint represents the allowable perturbation

$$\mathcal{D}^*_{d,\epsilon} := \{p_{Z|X,Y} \in \mathcal{P}(\mathcal{Z}|\mathcal{X},\mathcal{Y}) : \Pr[d(X,Z) \le \epsilon] = 1\}$$

Alternatively, can strengthen adversary by constraining only expected distortion

$$\mathcal{D}_{d,\epsilon} := \{p_{Z|X,Y} \in \mathcal{P}(\mathcal{Z}|\mathcal{X},\mathcal{Y}) : \mathbb{E}[d(X,Z)] \le \epsilon\}$$

More generally, we can consider closed, convex constraint set $\mathcal{D} \subset \mathcal{P}(\mathcal{X} \times \mathcal{Y})$

$$\min_{\theta} \max_{p_{X,Y} \in \mathcal{D}} \mathbb{E}[\ell(f_\theta(X), Y)]$$

MITSUBISHI ELECTRIC
Changes for the Better

Ideal Robust ML Equivalent to Privacy-Utility Tradeoff Problem

Consider *ideal* minimax solution over all classifiers (distributions) $q \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$

Theorem (Minimax Result)

For any finite sets $\mathcal{X}$ and $\mathcal{Y}$, and closed, convex $\mathcal{D} \subset \mathcal{P}(\mathcal{X}, \mathcal{Y})$, we have

$$\min_{q \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} \max_{p \in \mathcal{D}} \mathbb{E}[-\log q(Y|X)] = \max_{p \in \mathcal{D}} \min_{q \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} \mathbb{E}[-\log q(Y|X)]$$
$$= \max_{p \in \mathcal{D}} H(Y|X) =: h^* \leq \log |\mathcal{Y}|$$

where expectations and entropy are with respect to $(X, Y) \sim p$. Further, the solutions for $q \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ that solve the minimax (LHS) problem are given by

$$\bigcap_{p \in \mathcal{D}} \left\{ q \in \mathcal{P}(\mathcal{Y}|\mathcal{X}) : \mathbb{E}_{(X,Y) \sim p}[-\log q(Y|X)] \leq h^* \right\} \neq \varnothing.$$

RHS is a well-known, info-theoretic formulation of privacy-utility tradeoff

- Robust rule q^* (for LHS) must be consistent with $p^*_{Y|X}$ (from RHS optimum)
- Solving the max-entropy problem helps find minimax robust solution

MITSUBISHI ELECTRIC
Changes for the Better

Characterization of Robust Models

Corollary (Solution Set)

Under paradigm of above theorem, let $\mathcal{D}^* := \{p \in \mathcal{D} : H(Y|X) = h^*, (X,Y) \sim p\}$. *For all* $p^* \in \mathcal{D}^*$, *the corresponding terms of the solution set* $\bigcap_{p \in \mathcal{D}} Q(p)$ *are given by*

$$Q(p^*) := \{q \in \mathcal{P}(\mathcal{Y}|\mathcal{X}) : \mathbb{E}_{(X,Y)\sim p^*}[-\log q(Y|X)] \le h^*\}$$
$$= \{q \in \mathcal{P}(\mathcal{Y}|\mathcal{X}) : \forall (x,y), q(y|x)p^*(x) = p^*(x,y)\}.$$

Further, if

$$\bigcup_{p^* \in \mathcal{D}^*} \{x \in \mathcal{X} : p^*(x) > 0\} = \mathcal{X},$$

then the solution set contains exactly one point and is given by

$$\bigcap_{p^* \in \mathcal{D}^*} Q(p^*) = \bigcap_{p \in \mathcal{D}} Q(p).$$

If there exists $p^* \in \mathcal{D}^*$ with full support over $\mathcal{X}$ (in marginal P_X), then $q^* = p^*(y|x)$

Necessity of Stochastic Perturbation

Mixed (stochastic) strategies for adversary essential to the minimax equality

- No inherent disadvantage in playing first versus second

However, pure (deterministic) strategy adversaries at disadvantage when playing first

$$\min_{q \in \mathcal{P}(\mathcal{Y}|\mathcal{Z})} \mathbb{E}\left[\max_{\substack{Z \in \mathcal{X}: \\ d(X,Z) \leq \epsilon}} -\log q(Y|Z) \right] \geq \max_{\substack{g: \mathcal{X} \times \mathcal{Y} \to \mathcal{X} \\ d(X, g(X,Y)) \leq \epsilon}} \min_{q \in \mathcal{P}(\mathcal{Y}|\mathcal{Z})} \mathbb{E}\Big[-\log q\big(Y|g(X,Y)\big)\Big]$$

MITSUBISHI ELECTRIC
Changes for the Better

Necessity of Stochastic Perturbation

Mixed (stochastic) strategies for adversary essential to the minimax equality

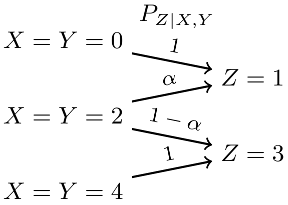
- No inherent disadvantage in playing first versus second

However, pure (deterministic) strategy adversaries at disadvantage when playing first

$$\min_{q \in \mathcal{P}(\mathcal{Y}|\mathcal{Z})} \mathbb{E}\left[\max_{\substack{Z \in \mathcal{X}: \\ d(X,Z) \leq \epsilon}} -\log q(Y|Z) \right] \geq \max_{\substack{g: \mathcal{X} \times \mathcal{Y} \to \mathcal{X} \\ d(X, g(X,Y)) \leq \epsilon}} \min_{q \in \mathcal{P}(\mathcal{Y}|\mathcal{Z})} \mathbb{E}\Big[-\log q\big(Y | g(X,Y)\big)\Big]$$

Example demonstrating strict inequality: $\mathcal{X} = \mathcal{Y} = \{0,1,2,3,4\}$, $X \sim$ Unif$\{0,2,4\}$, $X = Y$, and $d(X,Y) := |X - Z| \leq \epsilon = 1$

- Stochastic $P^*_{Z|X,Y} \Rightarrow \alpha = 0.5$, $\max H(Y|Z) = h_2(1/3)$
- Deterministic $g^* \Rightarrow \alpha = 0, 1$



Deterministic adversary: LHS (minimax) $h_2(1/3) > (2/3)\log(2)$ RHS (maximin)

Clean vs Robust Performance Tradeoffs

Theoretical analysis of "clean data penalty" for a robust model q^*

1. Ideal Bayes risk (of non-robust model on clean data): $H(Y|X)$
2. Loss for robust model on clean data: $H(Y|X) + \mathrm{KL}(p_{Y|X} \| q^* | p_X)$
3. Worst-case attack loss for robust model: $\max_{p_{X,Y} \in \mathcal{D}} H(Y|X)$

Note that $(1) \leq (2) \leq (3)$

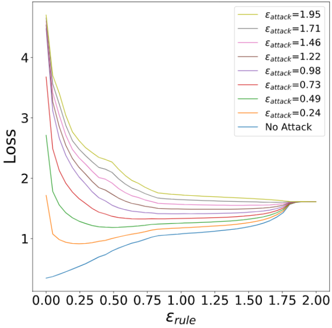
Clean vs Robust Performance Tradeoffs

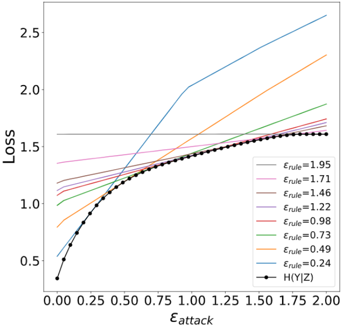
Theoretical analysis of "clean data penalty" for a robust model q^*

1. Ideal Bayes risk (of non-robust model on clean data): $H(Y|X)$
2. Loss for robust model on clean data: $H(Y|X) + \mathrm{KL}(p_{Y|X} \| q^* | p_X)$
3. Worst-case attack loss for robust model: $\max_{p_{X,Y} \in \mathcal{D}} H(Y|X)$

Note that $(1) \leq (2) \leq (3)$

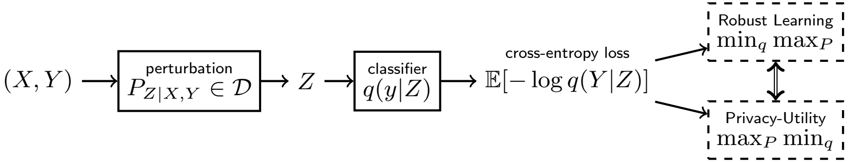
Mismatch between robust decision rule and attack strength leads to suboptimality





MITSUBISHI ELECTRIC
Changes for the Better

Conclusions and Further Work



Minimax result offers approach toward attaining robust models

- Solve max-entropy problem to find universal adversarial perturbation
- Optimal response to the universal adversary produces a robust model
- Considering stochastic adversaries necessary for saddle point
- Connections to privacy-utility theory help understand clean vs robust tradeoffs

See our extended paper on arXiv [2007.11693] for further details

- Generalization of main result to continuous alphabets
- Fixed-point characterization under Wasserstein ball constraints
- Ongoing investigation and application to robust learning methods