

NeurIPS2018 論文読み会

# "Adversarial vulnerability for any classifier"

Masanari Kimura

Ridge-i Inc., @machinery81

December 21, 2018

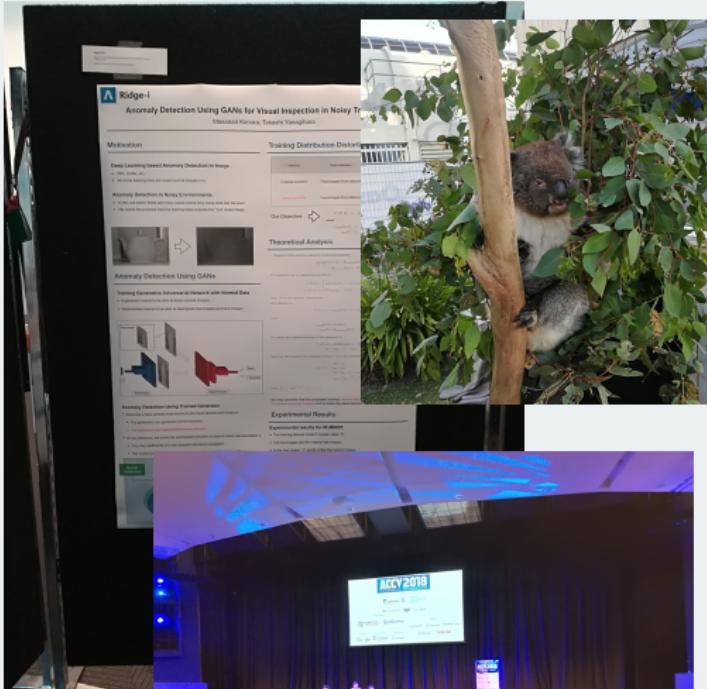
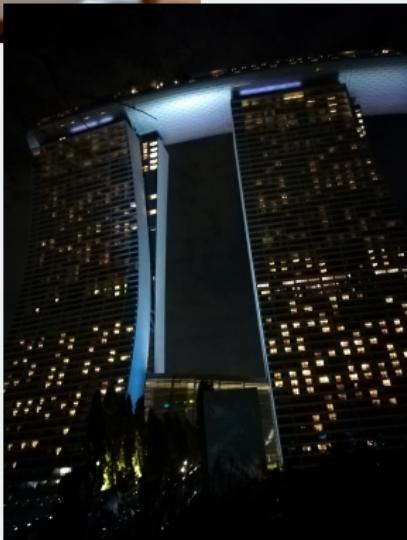
# ABOUT

# ABOUT

- Twitter やってます
- @machinery81
- C++, 離散構造, 生成モデル, 表現学習が好きです



# A TRAVEL NOTE OF 2018



# ABSTRACT

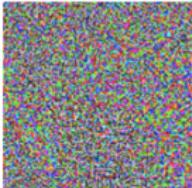
# Adversarial vulnerability for any classifier

- NeurIPS2018 採択論文 [1]
- Adversarial Examples についての論文
- 任意の分類器が達成可能な、Adversarial Examples に対する robustness のバウンドを導出した

# ADVERSARIAL ATTACKS

# WHAT IS ADVERSARIAL ATTACK?

- Classifier に対する攻撃手法 [2]
- 本来なら正しく分類できた画像に目視できないノイズをのせることで誤分類を誘発する

$$\begin{array}{ccc} \text{} & + .007 \times & \text{} \\ \text{$x$} & & \text{sign}(\nabla_x J(\theta, x, y)) \\ \text{"panda"} & & \text{"nematode"} \\ 57.7\% \text{ confidence} & & 8.2\% \text{ confidence} \end{array} = \begin{array}{c} \text{} \\ \text{$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$} \\ \text{"gibbon"} \\ 99.3 \% \text{ confidence} \end{array}$$

## THREAT OF ADVERSARIAL ATTACKS

e.g.

- 自動運転車に対する標識の誤検出
- 顔認証システムへの攻撃

...

...  
DNN の現実世界への適用において重大な課題.

# ADVERSARIAL ATTACKS AND DEFENCES

- 攻撃手法と防御手法のいたちごっこ
- 現時点で、完璧な防御手法は提案されていない

Methods	Year
Szegedy et al.[2]	2013
Goodfellow et al.[3]	2014
Papernot et al.[4]	2016
Dong et al.[5]	2017

Table 1: Attack Methods

Methods	Year
Papernot et al.[6]	2015
Papernot et al.[7]	2016
Tramer et al.[8]	2017
Athalye et al.[9]	2018

Table 2: Defense Methods

# ADVERSARIAL VULNERABILITY FOR ANY CLASSIFIER

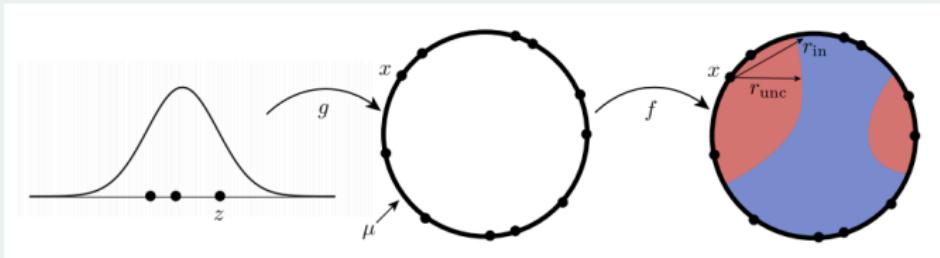
主張: 任意の分類器は Adversarial Attack に対する脆弱性をもつ

- 分類器が達成可能な Adversarial Attack に対するロバスト性のバウンドを示す
- 異なる複数の分類器の間で使いまわせる Adversarial Examples の存在を証明する

## ASSUMPTION

データは滑らかな生成モデルからマッピングされると仮定.

- 図の例では、正規分布からサンプリングされた  $z$  を円上にマッピング
- 分類器  $f$  はマッピングされたデータに対して分類面を引く (red or blue)



## SMOOTHNESS OF GENERATOR

生成モデルの滑らかさとは？

- 以下を満たすとき  $g(z)$  は十分滑らかであるとする
- 関数内の任意の 2 点間の値が  $\omega$  を超えないことを意味

$$\forall z, z' \in \mathcal{Z}, \quad \|g(z) - g(z')\| \leq \omega(\|z - z'\|_2) \quad (1)$$

## DEFINITION OF ROBUSTNESS

論文では、ロバスト性についての二つの定義を導入している。

### In-distribution robustness:

画像の潜在空間にノイズが与えられると仮定。

$$r_{in}(x) = \min_{r \in \mathbb{Z}} \|g(z + r) - x\| \text{ s.t. } f(g(z + r)) \neq f(x) \quad (2)$$

### Unconstrained robustness:

画像そのものにノイズが与えられると仮定。

$$r_{unc}(x) = \min_{r \in \chi} \|r\| \text{ s.t. } f(x + r) \neq f(x) \quad (3)$$

$r$  は与えるノイズ。直感的に、 $r_{unc}(x) \leq r_{in}(x)$ 。

## DEFINITION OF ROBUSTNESS

どちらも、”分類器が誤分類する最小のノイズ”を定義している。

- ・ノイズが十分大きければ分類器が誤分類するのは当たり前
- ・ノイズが小さすぎると誤分類させるのは難しい
- ・分類器が誤分類するギリギリのラインをロバスト性と定義

この論文のやりたいことは？

...

Adversarial Example に対する robustness のバウンドを求める

- 任意の分類器が常に脆弱性を持つことを示せる
- 我々が目指すべきロバスト性のベースラインがわかる

$r_{unc}(x) \leq r_{in}(x)$  から、

- in-distribution robustness  $r_{in}(x)$  の存在を示せれば、一般的に使われている  $r_{unc}(x)$  の上界の存在も示せる

# THEOREM 1. IN-DISTRIBUTION ROBUSTNESS

## THEOREM 1. IN-DISTRIBUTION ROBUSTNESS

任意の分類器  $f: \mathbb{R}^m \rightarrow \{1, \dots, K\}$  について,  $\eta$  より robustness が小さいデータが存在する割合は,

$$\mathbb{P}(r_{in}(x) \leq \eta) \geq \sum_{i=1}^K (\Phi(a_{\neq i} + \omega^{-1}(\eta)) - \Phi(a_{\neq i})) \quad (4)$$

ここで,  $\Phi$  は正規分布の累積分布関数. クラス分布が imbalanced でないとすると,

$$\mathbb{P}(r_{in}(x) \leq \eta) \geq 1 - \sqrt{\frac{\pi}{2}} e^{-\omega^{-1}(\eta)^2/2} \quad (5)$$

## THEOREM 1. IN-DISTRIBUTION ROBUSTNESS

式(4)について、 $g$ がリップシツ連続( $\Leftarrow$ 滑らか)とすると、  
連続率  $\omega^{-1}(\eta) = \eta/L$ .

$$\mathbb{P}(r_{in}(x) \leq \eta) \geq 1 - \sqrt{\frac{\pi}{2}} e^{-(\eta/L)^2/2} \quad (6)$$

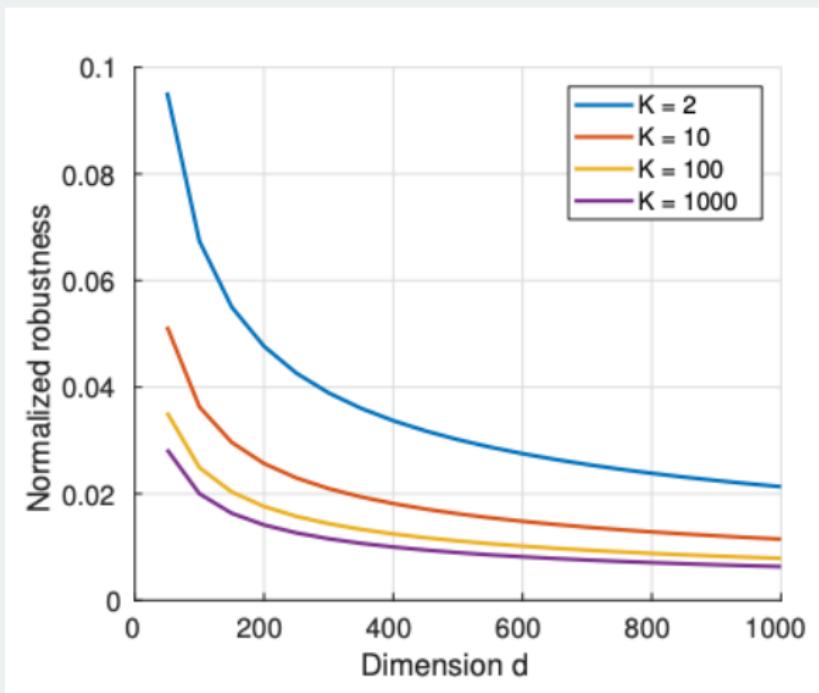
$\eta \propto L$ で、リップシツ定数  $L$  は関数の変化に対応するので、

- $g$  の傾きが小さいほど robustness は小さい
- 直感的には、データのバリエーションが多いほど robustness は小さくなる

## THEOREM 1. IN-DISTRIBUTION ROBUSTNESS

クラス数, 次元数と robustness との関係.

- クラス数が少ないほど robustness は大きい
- データの次元数が小さいほど robustness は大きい



# THEOREM 2. UNCONSTRAINED ROBUSTNESS

## THEOREM 2. UNCONSTRAINED ROBUSTNESS

$\tilde{f}$  を以下のように定義する:

$$\tilde{f}(x) = f(g(z^*)) \text{ with } z^* = \operatorname{argmin}_z \|g(z) - x\| \quad (7)$$

$f, g$  がどちらも同じ robustness を持つと仮定すると,  
 $\tilde{f}$  についての robustness は,  $r_{unc}(x) \geq \frac{1}{2}r_{in}(x)$

# THEOREM 3. TRANSFERABILITY OF PERTURBATIONS

## THEOREM 3. TRANSFERABILITY OF PERTURBATIONS

- 異なるモデル間で使いまわせる Adversarial Examples についての既存研究もいくつか存在
- データが滑らかな生成モデルがらマッピングされると仮定した時、こうした Adversarial Examples の存在は理論的に証明できる

## THEOREM 3. TRANSFERABILITY OF PERTURBATIONS

- $f, h$  をそれぞれ異なる分類器とする
- $\mathbb{P}(f \circ g(z) \neq h \circ g(z)) \leq \delta$  と仮定すると,

$$\begin{aligned} & \mathbb{P}\left\{\exists v : \|v\|_2 \leq \eta \text{ and } \begin{array}{l} f(g(z) + v) \neq f(g(z)) \\ h(g(z) + v) \neq h(g(z)) \end{array}\right\} \\ & \geq 1 - \sqrt{\frac{\pi}{2}} e^{-\omega^{-1}(\eta)^2/2} - 2\delta = 1 - \epsilon \end{aligned}$$

言い換えると,  $f$  と  $h$  を両方騙せるノイズ  $v$  が存在する確率は,  
 $1 - \epsilon$  より大きい。

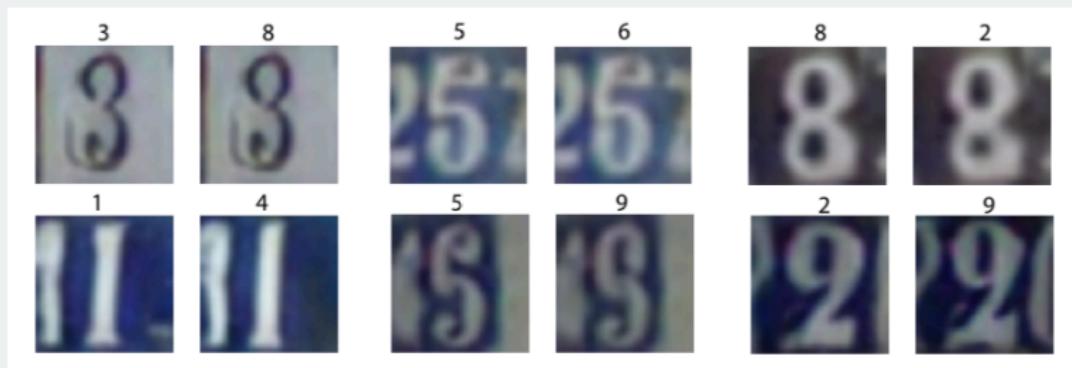
$\epsilon$  は“分類器を騙せないリスク”を意味し, 十分小さな値になる。

# EXPERIMENTAL EVALUATION

## EXPERIMENTAL EVALUATION

- SVHN データセットにおける実験.
- 分類器は ResNet-18
- 左が元画像, 右が Adversarial Example

Figure 1: Illustration of generated images.



## EXPERIMENTAL EVALUATION

- SVHN データセットにおける実験.
- 既存のネットワークアーキテクチャについて robustness を評価.
- $r_{unc}(x) \leq r_{in}(x)$  を満たしている

Figure 2: Experiments on SVHN dataset.

	Upper bound on robustness	2-Layer LeNet	ResNet-18	ResNet-101
Error rate	-	11%	4.8%	4.2 %
Robustness in the $\mathcal{Z}$ -space	$16 \times 10^{-3}$	$6.1 \times 10^{-3}$	$6.1 \times 10^{-3}$	$6.6 \times 10^{-3}$
In-distribution robustness	$36 \times 10^{-2}$	$3.3 \times 10^{-2}$	$3.1 \times 10^{-2}$	$3.1 \times 10^{-2}$
Unconstrained robustness	$36 \times 10^{-2}$	$0.39 \times 10^{-2}$	$1.1 \times 10^{-2}$	$1.4 \times 10^{-2}$

## EXPERIMENTAL EVALUATION

- CIFAR-10 データセットにおける実験.
- 既存のネットワークアーキテクチャについて robustness を評価.
- $r_{unc}(x) \leq r_{in}(x)$  を満たしている

Figure 3: Experiments on CIFAR-10 (same setting as in Table 1).

	Upper bound on robustness	VGG [40]	Wide ResNet [41]	Wide ResNet + Adv. training [10, 15]
Error rate	-	5.5%	3.9%	16.0%
Robustness in the $\mathcal{Z}$ -space	0.016	$2.5 \times 10^{-3}$	$3.0 \times 10^{-3}$	$3.6 \times 10^{-3}$
In-distribution robustness	0.10	$4.8 \times 10^{-3}$	$5.9 \times 10^{-3}$	$8.3 \times 10^{-3}$
Unconstrained robustness	0.10	$0.23 \times 10^{-3}$	$0.20 \times 10^{-3}$	$2.0 \times 10^{-3}$

## CONCLUSION & DISCUSSION

- 分類器には必ず Adversarial Attack に対する脆弱性が存在することを証明
- 全ての分類器が超えられない、Adversarial Attacks に対する robustness のベースラインを導出

# REFERENCES

# REFERENCES

-  Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi.  
Adversarial vulnerability for any classifier.  
In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pp. 1186–1195. Curran Associates, Inc., 2018.
-  Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus.  
Intriguing properties of neural networks.  
*arXiv preprint arXiv:1312.6199*, 2013.
-  Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy.  
Explaining and harnessing adversarial examples (2014).  
*arXiv preprint arXiv:1412.6572*.
-  Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami.  
The limitations of deep learning in adversarial settings.  
In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pp. 372–387. IEEE, 2016.
-  Yingpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Xiaolin Hu, Jianguo Li, and Jun Zhu.  
Boosting adversarial attacks with momentum.  
*arXiv preprint arXiv:1710.06081*, 2017.
-  Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami.  
Distillation as a defense to adversarial perturbations against deep neural networks.  
In *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 582–597. IEEE, 2016.
-  Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman.  
Towards the science of security and privacy in machine learning.  
*arXiv preprint arXiv:1611.03814*, 2016.
-  Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel.  
Ensemble adversarial training: Attacks and defenses.  
*arXiv preprint arXiv:1705.07204*, 2017.