紹介論文

Adversarial Examples Are Not Bugs, They Are Features

Ilyas, Andrew and Santurkar, Shibani and Tsipras, Dimitris and Engstrom, Logan and Tran, Brandon and Madry, Aleksander

In *Proceedings of Advances in Neural Information Processing Systems 32 (NeurIPS2019)*

1 What is claim on this paper

- AE (Adversarial Example) は non-robust feature (features (derived from patterns in the data distribution) that are highly predictive, yet brittle and (thus) incomprehensible to humans) の存在によって引き起こされるものである
- "Adversarial vulnerability is a direct result of sensitivity to well-generalizing features in the data"
  - 単純に分類精度を最大化させる分類器を学習する場合, 人間には理解できないものも含んだ利用可能な全ての signal を使う傾向がある (耳や目などの人間では知覚できる signal 以外も用いる).
  - このような人間によって理解できないが, 精度向上に繋がる "non-robust" feature が存在する. non-robust feature を悪用することで敵対的な摂動 (攻撃) に繋がると推測している.
  - この仮説は, adversarial transferabillity (あるモデルに対して計算された摂動が他の独立して訓練されたモデルに転移すること) の可能性を示唆している. どのモデルも non-robust feature を学習する可能性がある (予測には有用なため), それを用いて摂動を与えることで, 転移させることができる. つまり, モデルの bug ではなく, データに内在する feature であると考えられる.
  - 逆に, データセットから non-robust feature を取り除けば AE に対して堅牢なモデルを作成できるのではないのかと考えられる.

2 Experiment result and insight

- robust feature のみのデータセットを作成して, AE に対する堅牢性を評価する. → 敵対的な脆弱性は必ずしも学習手法に起因しているのではなく, データセットにも起因していることをがわかる.
- 入力が元のデータとほぼ同じであるが, 正しくラベル付けされてされていないデータを作成する. 実際には, 敵対的な摂動を介してのみラベル付けされている (hence utilize only non-robust feature). このような non-robust feature を用いて学習を行い分類精度を評価する. 人 → 人間によって予測可能な情報がないにも関わらず, このデータセットで学習した場合, テストデータセット (摂動が加えられていない正常なデータ) で精度が向上する. これは正しい入力の分類に役立つデータの特徴を反転させることで敵対的な摂動が発生する可能性がある. (?) (his demonstrates that adversarial perturbations can arise from flipping features in the data that are useful for classification of correct inputs (hence not being purely aberrations).)

3 Robust Features Model

3.1 Setup

2 値分類を考える ($(x, y) \in \mathcal{X} \times \{\pm 1\}$ are sampled from a distribution $\mathcal{D}$). 学習器 $\mathcal{C}$ を学習する. 特徴量を入力空間から実数への写像と定義して, 全ての関数集合を $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$ と定義する. 簡単のため, $\mathbb{E}_{(x,y)\sim\mathcal{D}}[f(x)] = 0, \mathbb{E}_{(x,y)\sim\mathcal{D}}[f(x)^2] = 1$ とする.

3.2 Useful, robust, and non-robust features

以下のように定義する.

- ρ-**useful features** : 分布 $\mathcal{D}$ が与えられたもとで, 以下を満たす場合, 特徴量 f は ρ-useful ($\rho > 0$) であると呼ぶ. また, 分布 $\mathcal{D}$ のもとで, f が ρ-useful である最大の ρ を $\rho_{\mathcal{D}}(f)$ とする.

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[y \cdot f(x)] \geq \rho \tag{1}$$

- γ-**robustly useful features** : 特徴量 f は ρ-useful ($\rho_{\mathcal{D}}(f) > 0$) であると仮定する. 以下を満たす場合, f は γ-robust feature (formally a γ-robustly useful feature for $\gamma > 0$) と呼ぶ.

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\inf_{\delta\in\Delta_{(x)}} y \cdot f(x+\delta)\right] \geq \gamma \tag{2}$$

- **Useful, non-robust features** : 特徴量 f は rho が 0 を離れた値をとる ρ-useful であるが, 任意の $\gamma \geq 0$ で γ-robust feature を満たさない場合, useful, non-robust feature と呼ぶ.

3.3 Classification

特徴量の集合 $F \subseteq \mathcal{F}$, 重みベクトル w, バイアス b として, 入力 x に対して学習器は以下のように予測する.

$$C(x) = \mathrm{sgn}\left(b + \sum_{f\in F} w_f \cdot f(x)\right) \tag{3}$$

3.4 Standard training

分類器の学習は以下の損失の最小化を行う. このような最小化は robust features と non-robust features の区別を行わず, usefulness のみに着目したものである.

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}_\theta(x, y)] = -\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[y\left(b + \sum_{f\in F} w_f \cdot f(x)\right)\right] \tag{4}$$

GETROBUSTDATASET(D)

1. $C_R \leftarrow$ ADVERSARIALTRAINING(D)
   $g_R \leftarrow$ mapping learned by C_R from the input to the representation layer
2. $D_R \leftarrow \{\}$
3. For $(x, y) \in D$
   $x' \sim D$
   $x_R \leftarrow \arg\min_{z \in [0,1]^d} \|g_R(z) - g_R(x)\|_2$
   # Solved using ℓ_2-PGD starting from x'
   $D_R \leftarrow D_R \bigcup \{(x_R, y)\}$
4. Return D_R

Figure1　Algorithm to construct a “robust” dataset, by restricting to features used by a robust model

3.5 Robust training

non-robust feature は有用な特徴であるが, 真のラベルを anti-correlated させうるので, robust feature のみを学習するために, adversarial loss function を利用した学習を行う.

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\delta\in\Delta_{(x)}} \mathcal{L}_\theta(x+\delta, y)\right] \tag{5}$$

4 Finding Robust (and Non-Robust) Features

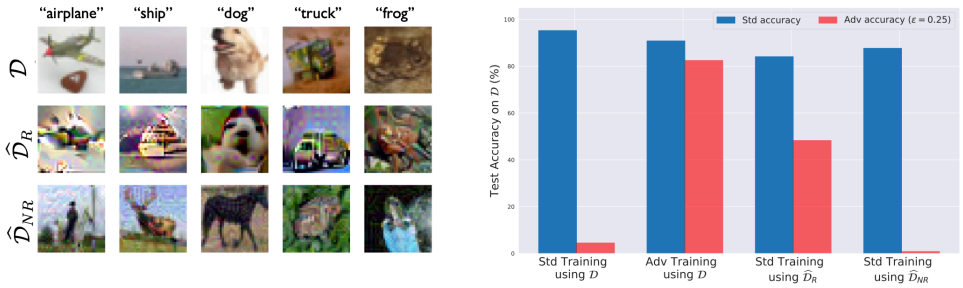
4.1 Disentangling robust and non-robust features

現状では, データセットから robust featue を抜き出す方法が存在しない. そこで robust training で得られた学習器 C　(robust model) を使用して, robust feature を作成する. つまり, 以下を満たすような分布 $\hat{D}_R$ を作成する.

$$\mathbb{E}_{(x,y)\sim\hat{\mathcal{D}}_\mathcal{R}}[f(x)\cdot y] = \begin{cases} \mathbb{E}_{(x,y)\sim\mathcal{D}}[f(x)\cdot y] & (f \in F_C) \\ 0 & (\text{otherwise}) \end{cases} \tag{6}$$

where F_C again represents the set of features utilized by C. つまり, robust model から得られる feature (robust feature) のみで構成されるデータセットを作成するということである.　これの逆 (robust model で得られない特徴量 → standard training から得られる特徴量 → non-robust feature) のデータセットを $\hat{D}_{NR}$ とする (ここで non-robust と言っているが, non-robust のみであるとは限らないことに注意). 具体的には, Figure 1 の手続きに従ってデータセットを作成する.

上記の手続きで作成されたデータセット (a) と, 得られたデータセットで Standard training (Std training) または, Robust training (Adversarial training, Adv training) を行い, 通常のテストデータセットまたは, 敵対的摂動を加えたテストデータセットで精度を比較した図 (b) を Figure 2 に示す.



(a) (b)

Figure 2: **(a)**: Random samples from our variants of the CIFAR-10 [Kri09] training set: the original training set; the *robust training set* $\widehat{\mathcal{D}}_R$, restricted to features used by a robust model; and the *non-robust training set* $\widehat{\mathcal{D}}_{NR}$, restricted to features relevant to a standard model (labels appear incorrect to humans). **(b)**: Standard and robust accuracy on the CIFAR-10 test set ($\mathcal{D}$) for models trained with: (i) standard training (on $\mathcal{D}$) ; (ii) standard training on $\widehat{\mathcal{D}}_{NR}$; (iii) adversarial training (on $\mathcal{D}$); and (iv) standard training on $\widehat{\mathcal{D}}_R$. Models trained on $\widehat{\mathcal{D}}_R$ and $\widehat{\mathcal{D}}_{NR}$ reflect the original models used to create them: notably, standard training on $\widehat{\mathcal{D}}_R$ yields nontrivial robust accuracy. Results for Restricted-ImageNet [Tsi+19] are in D.8 Figure 12.

Figure2　Robust dataset and result

4.2　Non-robust features suffice for standard classification

次に non-robust feature のみを含むデータを作成する．Figure 3 の手続きに従って作成する．元のデータセットから一様ランダム (uniformly at random) または決定的 (deterministically) にラベルを書き換える $(y \to t)$. その後, 変更したラベルの損失が小さくなるように摂動を加える $(x \to \boldsymbol{x}_{adv})$.

加えられた摂動が小さいとき, x_{adv} は y との相関はまだ存在している．一方で, non-robust feature は t に相関したものとなっている．

t がランダムに決められた場合, robust feature は元々ラベル t と相関が無いため, 摂動が加えられても, 相関は低い．よって, データセット $\hat{D}_{rand}$ を作成することを目標とする．

$$\mathbb{E}_{(x,y)\sim\hat{\mathcal{D}}_{rand}}[f(x)\cdot y] = \begin{cases} > 0 & if\ f\ non-robustly\ useful\ under\ \mathcal{D} \\ \simeq 0 & (\text{otherwise}) \end{cases} \tag{7}$$

一方で, t が y に基づいて決定的に決められた場合, robust feature は t と離れている．実際, non-robust feature は t に相関するが, robust feature は y に相関する．robust feature は訓練データセットを予測するには予測力があるが, テストデータセットを予測するときは精度を悪化させる．つまり, データセット $\hat{D}_{det}$ を作成することを目標とする．

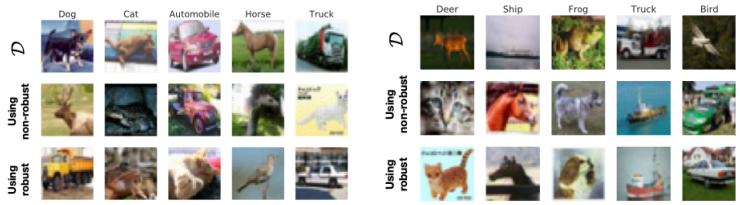
$$\mathbb{E}_{(x,y)\sim\hat{\mathcal{D}}_{dat}}[f(x)\cdot y] = \begin{cases} > 0 & if\ f\ non-robustly\ useful\ under\ \mathcal{D} \\ < 0 & if\ f\ robustly\ useful\ under\ \mathcal{D} \\ \in \mathbb{R} & if\ f\ not\ useful\ under\ \mathcal{D} \end{cases} \tag{8}$$

作成されたデータセットを Figure 4 に示す．元の画像に摂動を加えてラベルを変更することでデータセットを作成するため, 人間の目から見ると誤ったラベリングがされているように思われて, 人間には理解できないデータセットとなる．

GETNONROBUSTDATASET(D, ε)

1. $D_{NR} \leftarrow \{\}$
2. $C \leftarrow$ STANDARDTRAINING(D)
3. For $(x, y) \in D$
   $t \overset{\text{uar}}{\sim} [C]$ # or $t \leftarrow (y+1) \mod C$
   $x_{NR} \leftarrow \min_{||x'-x|| \leq \varepsilon} L_C(x', t)$ # Solved using ℓ_2 PGD
   $D_{NR} \leftarrow D_{NR} \bigcup \{(x_{NR}, t)\}$
4. Return D_{NR}

Figure3　Algorithm to construct a dataset where input-label association is based entirely on non-robust features



(a) $\widehat{\mathcal{D}}_{rand}$　　(b) $\widehat{\mathcal{D}}_{det}$

Figure4　Non robust dataset

4.3　Transferability can arise from non-robust features

AE が non-robust feature の存在によって引き起こされるものならば, 同じ non-robust feature を学習することで, AE の転移を引き起こすことが可能になる.

5 つの異なるモデルを, standard training された ResNet-50 を用いて作られた non-robust のみのデータセット (前章の作成方法で作成) で学習して, 精度と transfer success rate を評価した (Figure 5). 精度と transfer success rate は相関しており, 精度が高い (non-robust features をよく学習している) モデルほど生成元の ResNet-50 から生成された adversarial examples に騙されやすいという結果が得られた.

4.4　Conclusion

解釈性の観点から学習モデルが non-robust feature に依存している限り, 人間にとって意味のありモデルに忠実な説明を期待することはできない. 堅牢で解釈可能なモデルを実現するには人間の事前知識を訓練プロセスに明示的に取り入れる必要がある.

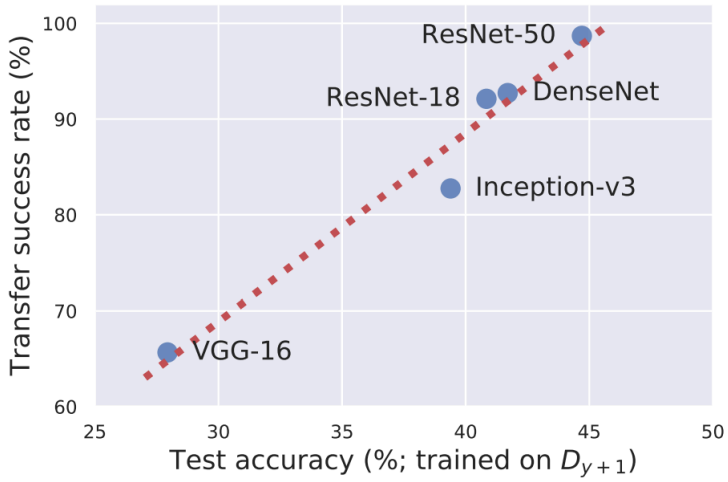


Figure5 Transfer rate of adversarial examples from a ResNet-50 to different architectures alongside test set performance of these architecture