On an adversarial example attack by partially monitoring and tampering input features to machine learning models - its possibilities and countermeasures -

角田　有紀子
大阪大学 大学院情報科学研究科 村田研究室

研究背景

- **機械学習の普及**
  - 工場や農場での異常検出など、さまざまな分野で利用
  - 機械学習ベースのアプリケーションも注目
    - スマートヘルスケアやスマートホームなど
- **機械学習モデルへの攻撃が懸念**
  - モデルを誤判別させる小さなノイズを載せる敵対的サンプルなど
    - 敵対的サンプルの例：人がパンダと判別する画像をテナガザルと誤判別 [1]
    - 「止まれ」の交通標識を「制限時速 45 マイル」と誤判別 [2]
  - 対策手法も検討されているが、根本的な解決策は未発見

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarialexamples,"arXiv preprint arXiv:1412.6572, 2014
[2] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash,T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual clas-sification," inProceedings of the IEEE Conference on Computer Vision and PatternRecognition, 2018, pp. 1625–1634

複数センサー利用のシステムにおける攻撃リスク

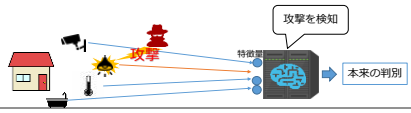
- **複数センサー利用のシステムの場合において、センサーを用いた攻撃でも誤判別させられる可能性がある場合、敵対的サンプルのリスクは高くなる**
  - 攻撃者は脆弱性のあるセンサーを利用して攻撃する可能性

研究目的

- 攻撃者が入力特徴量の一部を監視・改ざん可能な場合における攻撃の可能性の検証
- 上記の攻撃への対策を提案

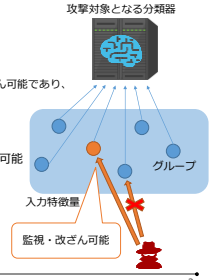
アプローチ

攻撃：入力特徴量の一部を入力して攻撃生成できる機械学習モデルを作成
対策：攻撃された特徴量を見つけて、除外して判別



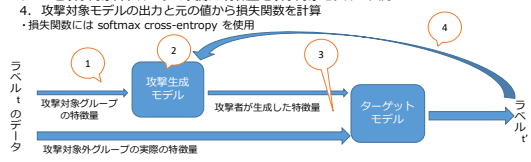
本研究で対象とする攻撃の定義

- **攻撃対象**
  - 入力特徴に対応するラベルを出力する分類器
- **攻撃対象の分類器への入力特徴量**
  - 入力特徴量は複数のグループに分けることが可能
    - 各グループは、監視と改ざんが可能な特徴量を含む
    - 例：センサーの特徴量は、センサーをハッキングすることで改ざん可能であり、その場合、センサーの特徴量は同じグループに属する
- **攻撃者の能力**
  - 攻撃対象モデルがどのような分類器か把握可能
    - ネットワークの構造、パラメータ、トレーニングデータを含む
  - 攻撃生成時には 1 つのグループの特徴量のみを監視・改ざん可能
    - 監視できるグループ外の特徴量は取得不可
- **攻撃者の目標**
  - 標的型攻撃の成功 (以降、攻撃 = 標的型攻撃とする)
    - 攻撃者が指定したラベルを攻撃対象が出力



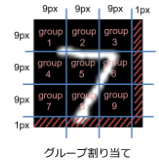
攻撃生成方法

- **攻撃生成モデルを使用して攻撃データを生成**
  - 攻撃生成モデルはトレーニングデータを使用して学習
- **攻撃生成モデルの学習手順**
  1. 攻撃対象グループの特徴量を攻撃生成モデルへ入力
  2. 攻撃生成モデルで攻撃対象グループの特徴量に対応する特徴量を生成
  3. 2 と攻撃対象外グループの実際の特徴量を攻撃対象モデルへ入力
  4. 攻撃対象モデルの出力と元の値から損失関数を計算
     - 損失関数には softmax cross-entropy を使用



攻撃の評価方法

- **データセット**
  - MNIST (28 ピクセル × 28 ピクセル)
    - 機械学習の分野で最も有名な手書き数字の画像データセット
  - 9 つのブロックに分割してグループを設定
- **攻撃対象**
  - 畳み込みニューラルネットワーク
- **攻撃の評価指標**
  - 成功率 = $\frac{\text{ターゲットラベルに分類された画像の数}}{\text{生成された画像の数}}$



攻撃の評価結果

- **24 個のパターンにおいて成功率が 0.7 以上**
  - うち、13 個のパターンにおいて成功率が 0.9 以上
  - グループ 5 を攻撃したときの成功率は特に高い
- **“1” を “7” と誤判別させる攻撃では特に高い攻撃成功率**

“1” を “7” と誤判別させる攻撃の
各グループに対する攻撃例

0.007	0.974	0.000
0.006	0.989	0.199
0.004	0.208	0.008

左図の攻撃成功率

攻撃対策

- **攻撃された特徴量を見つけて、除外して判別**
  - 判別結果への影響量に着目して攻撃の可能性を検知し、攻撃前ラベル候補と元の出力ラベルを比較することにより、攻撃によって誤判別させられたか判定
  - さらに、攻撃されていない場合の判別結果の候補を提示
- **攻撃検知の手順**

1. 判別結果に大きな影響を与えるグループを選出
影響量を求める手段として SmoothGrad [3] を使用し、
得られる値の合計値が閾値を超えた場合にそのグループを選出

SmoothGrad：
深層学習の判別理由を明確にする方法の 1 つ

2. 1 のグループの特徴量を使用せずに各ラベルに対する確率を取得
判別の際、新たなモデル (欠損対応モデル) を導入

3. 元の結果と 2 の結果を比較し、攻撃の有無を判定
a. 2 において、元の出力ラベルの確率が低い場合：
攻撃ありとし、2 の出力を攻撃前ラベル候補とする
b. a 以外について、他のラベルの確率に高いものが含まれる場合：
攻撃の可能性ありとし、2 の出力を攻撃前ラベル候補とする
c. それ以外：攻撃なし

[3] D. Smilkov, N. Thorat, B. Kim, F. Vi ́egas, and M. Wattenberg, “Smoothgrad: re-moving noise by adding noise,” 2017.

欠損対応モデル

- **入力特徴量に攻撃データが含まれていても攻撃前ラベルの推測が可能な機械学習モデル**

- **構造**
  - グループに対応する特徴量を入力
  - 各ラベルに対する確率を出力
    - ラベル候補が複数になることを許可
    - Sigmoid 関数を活性化関数として使用
- **学習方法**
  - 各グループの特徴量を欠損させて学習
    - 欠損のさせ方はグループ単位
    - 除外された特徴量は 0 に設定
  - 教師ラベルを含まない出力へのペナルティが大きくなるような誤差関数を使用
  - 攻撃対象のトレーニングデータと同じものを使用して学習

中間層 (j 番目の層の i 番目のノードの出力)

$$o_{i,j} = a\left(\sum_{k \in C_{i,j}} \frac{1}{1-N_{C_{i,j}}^{excluded}} \omega_{j-1,k,i} o_{k,j-1} + b_{i,j}\right)$$

$a()$: 活性化関数
C: ノードのセット
$N_{C_{i,j}}^{excluded}$: 除外された特徴量の数
ω, b: 重み

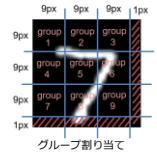
誤差関数

$$L(Y,T) = -\sum(\omega(t_i \log y_i + (1-t_i)\log(1-y_i)))$$

Y: モデルの出力ラベル　　T: 教師ラベル
y_i: Y の i 番目の要素　　t_i: T の i 番目の要素
$\omega(t)$: $\omega(0) \ll \omega(1)$となるよう設定された重み

対策手法の評価方法

- **データセット**
  - MNIST (28 ピクセル × 28 ピクセル)
  - 9 つのブロックに分割してグループを設定
- **攻撃対象**
  - 畳み込みニューラルネットワーク
- **評価指標**
  - 検知率 $= \dfrac{\text{攻撃断定された画像の数}}{\text{生成された攻撃数}}$
  - 誤検知率 $= \dfrac{\text{正常と判別された画像の数}}{\text{生成された攻撃数}}$



攻撃対策手法の評価結果

- **攻撃の検知率は88%で、誤検知率は1%**
- **オリジナルデータの40%は攻撃の可能性有と分類**
  - 今後の課題の一つ
- **攻撃前ラベル推測の成功率は96%**
  - 240件中230件で正しいラベルを含む候補を出力
  - そのうち、ラベルが完全一致したのは132件

攻撃検出結果

	攻撃断定	攻撃の可能性有	正常	合計
攻撃データ	212	26	2	240
オリジナルデータ	1	40	59	100

攻撃前ラベル推測結果

	完全一致	一部一致	失敗
攻撃データ	132	98	10

まとめと今後の課題

- **本研究のまとめ**
  - 攻撃者が特徴量の一部を監視・改ざんできる場合に可能な攻撃方法を提示
    - 出力は、攻撃対象の誤った決定を引き起こす特徴量
    - 高い確率で攻撃成功する組み合わせが存在
  - 上記の攻撃に対する対策を提案し、その有効性を実証
    - 欠損対応モデルで攻撃前ラベルを出力
    - 元のモデルと欠損対応モデルの出力を比較し、攻撃を検出
- **今後の課題**
  - 欠損対応モデルの改善
  - センサーベースの機械学習モデルでの評価