講演番号：53

Adversarial Training の考察に基づく Adversarial Examples への耐性の向上

D－2

Improving Adversarial Robustness Based on Adversarial Training Consideration

小宮山 亮太† 服部 元信†

Ryota KOMIYAMA† Motonobu HATTORI†

† 山梨大学 大学院医工農学総合教育部

† Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences, University of Yamanashi

1. はじめに

近年ニューラルネットワークの進歩に伴い，様々なタスクで高性能を記録している．しかし高性能なモデルであっても Adversarial Examples と呼ばれる画像によって誤認識を引き起こすことが可能である[1]．本稿では，Adversarial Examples が入力されたときの中間ニューロンの活動に着目した手法によって，耐性を向上できることを報告する．

2. Adversarial Examples

Adversarial Examples(AE)とは通常画像に些細なノイズを付加して作成される画像である．図 1 に例を示す．学習後のネットワークによって，図1左の画像は正しく猫と認識される一方，右の画像はパンダと誤認識される．このように人間にはほとんど知覚できない差異であるのも関わらずネットワークが誤認識を引き起こす．

図 1. Adversarial Examples の例

AEが発生する原因として最も有力な仮説は，入力画像に耐性のない特徴が混入しているというものである[2]．しかし入力画像から耐性のない特徴を取り除いたとしても，一定の耐性は獲得できるものの依然として誤認識を引き起こす．そこで現在最も耐性を得ることができる従来手法 Adversarial Training がどのように耐性を獲得しているかを考察することで，さらなる耐性獲得を行う．

3. 簡易データによる考察

式(1, 2)に示すデータセットをネットワークに学習させる．

$$x_1, \dots, x_r = \begin{cases} +y \\ -y \end{cases} \dots (1) \qquad x_{r+1}, \dots, x_{r+d} \sim \mathcal{N}\left(\frac{3y}{\sqrt{d}}, 1\right) \dots (2)$$

ただしyは教師ラベル，r, dはそれぞれの特徴の次元数を表している．式(1)は耐性のある特徴を表現し，$+y$を選択する確率pによってデータセットの難易度を調節する．式(2)は耐性のない特徴を表現し，すべて足し合わせることで 99%以上の分類性能を達成可能である．

このデータセットを 3 層のニューラルネットワークに学習させ中間層を観察する．Adversarial Training は通常画像と AE を同じ教師ラベルとしてネットワークに学習する手法である．しかしニューロンを観察すると出力を同一視する学習を行っているにもかかわらず，中間層では区別できる情報を学習していることがわかった．そのため Adversarial Training をしつつ，中間層をより区別できるように学習することで，さらなる耐性を獲得可能であると考えた．

4. 提案手法

3.の考察から，区別可能な情報を学習するために．識別器に加え検出器を接続して Adversarial Training を行う手法を提案する．この手法では中間層の情報を検出器の入力となるようにネットワークを接続し，マルチタスク学習を行う．さらに識別器で誤認識した AE を重点的に学習できるように，検出器で通常画像と正しく識別した AE を通常画像として，誤認識した AE を AE として学習した．検出器はクラス数分だけ出力ニューロンがある構造であり，学習には識別器と検出器をどちらも考慮した AE を使用した[3]．識別器の学習および評価には PGD と呼ばれる手法を利用した[4]．

5. 実験結果

Adversarial Training による耐性，提案手法による耐性を表 1 に示す．付加したノイズは 0.1，使用した学習データは Fashion-MNIST である．結果は高い方が良い．考察通りに従来手法より有意に優れた結果を得ることができた．

表 1. 実験結果 (%)

	通常精度	Adv. 精度 (0.1)
Adv. Training	88.84	75.80
提案手法	**89.17**	**76.31**

6. 今後の課題

検出器の学習を工夫してさらなる耐性の向上を図る．

参考文献

[1] C. Szegedy, *et al.*, Intriguing properties of neural networks, ICML, 2014.

[2] D. Tsipras, *et al.*, Robustness May Be at Odds with Accuracy, ICML, 2019.

[3] N. Carlini, *et al.*, Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods, AIsec, 2017.

[4] A. Madry, *et al.*, Towards Deep Learning Models Resistant to Adversarial Attacks, ICML, 2017.