L-012

データ多様体の埋め込み幾何学に基づいた敵対的サンプルの検知手法
Detecting Adversarial Examples Based on Embedding Geometry of Data Manifolds

久重 広樹† 田崎 元† 趙 晋輝†
Hiroki Hisashige Hajime Tasaki Jinhui Chao

1. はじめに

深層ニューラルネットワーク（DNN）は様々な分野で活用され，特に画像認識はセキュリティ分野への応用が進んでいる．その一方で，人間には知覚困難なほどに小さなノイズを加えることで意図的に誤分類を引き起こす敵対的サンプルという攻撃が発見されており，これは DNN が脆弱性を持つことを意味している．そこで産業技術総合研究所が発行している機械学習に関するマネジメントガイドライン[1]や，EU のサイバーセキュリティ機関である ENISA が発行している自動運転技術の AI に関するレポート[2]に，敵対的サンプルとその対策についての記載があり，DNN の脆弱性は広く問題視されている．しかし，敵対的サンプルの発生メカニズムが解明されていないために，既存対策手法の効果は限定的であった．本研究では，多様体仮説のもとでデータ多様体の埋め込み幾何学に基づき，攻撃入力の特徴であるデータ多様体の接空間の直交補空間成分を利用して攻撃入力を検知する新たな手法を提案する．また，手書き文字画像から生成した攻撃画像を用いて提案手法の性能評価を行う．

2. 敵対的攻撃

敵対的攻撃とは，正常画像に対して人間には知覚することが困難なほど小さなノイズを加えた画像を用いて，ニューラルネットワークに意図的に誤分類させる攻撃である．このときに用いられる攻撃画像を敵対的サンプルと呼ぶ．

2.1 攻撃手法

2014 年に Szegedy らが敵対的攻撃を発見したことをきっかけに[3]，これまでに敵対的サンプルの様々な生成手法が発見されており， それに対して数多くの対策手法も提案されている[4][5]．対策手法は攻撃データに対して正規の予測を得ることを目的とする防御手法と，攻撃データの入力を未然に防ぐことを目的とした検知手法に分かれるが，本稿では検知手法について議論する．

2.2 検知手法

検知手法とは，入力データが正常データかそうでないかを判別する手法である．攻撃データをニューラルネットワークに入力された場合に，予測は行わず破棄するなどの運用が想定される．これにより防御手法の問題であった正常入力に対する分類精度の低下が起こらないと言うメリットがある．敵対的サンプルを検知する様々な研究が行われており[6]，特に Roth らが文献[7]で提案したランダムノイズを用いた検知手法がある．この手法は入力画像と，入力画像に対してランダムノイズを加えた画像に対するニューラルネットワークの出力に，特徴的な差が生じることに着目して，正常画像と攻撃画像を判別する．

†中央大学大学院理工学研究科
Graduate School of Science and Engineering, Chuo University

3. 先行研究

近年，田崎らによって敵対的サンプルの発生メカニズムが解明された[8]．画像データは画素数を次元とする高次元ベクトルで表すことができる．一方でそのデータ集合は多様体仮説のもとで低次元の多様体構造を持つことが知られている．このデータ集合の多様体をデータ多様体と呼び，データ多様体の局所近傍は接空間というアフィン空間で表すことができる．

n次元空間S内のデータ多様体M上にベクトル$\boldsymbol{x}$がある時，ニューラルネットワークの各ノードの出力は，重み$\boldsymbol{w}$とバイアスθ，活性化関数fを用いて$f(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\theta)$と表せるが，$n+1$次元空間$\mathcal{S}$に埋め込んだデータ多様体$\mathcal{M}=\{\boldsymbol{x}=(\boldsymbol{x}^{\mathrm{T}},1)^{\mathrm{T}}|\boldsymbol{x}\in M\}$を考え，重みを$\boldsymbol{w}=(\boldsymbol{w}^{\mathrm{T}},\theta)^{\mathrm{T}}$とすると，$f(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})$で計算が可能になる．ここで$\mathcal{M}$上のベクトル$\boldsymbol{x}$における接空間を$T_{\boldsymbol{x}}\mathcal{M}$，これに対する直交補空間を$T_{\boldsymbol{x}}^{\perp}\mathcal{M}$として直交分解$T_{\boldsymbol{x}}\mathcal{S}=T_{\boldsymbol{x}}\mathcal{M}\bigoplus T_{\boldsymbol{x}}^{\perp}\mathcal{M}$が成り立つ．同様にして$\boldsymbol{x}$は$\boldsymbol{x}=\boldsymbol{x}_{\mathcal{M}}+\boldsymbol{x}_{\mathcal{M}}^{\perp}$，$\boldsymbol{w}$は$\boldsymbol{w}=\boldsymbol{w}_{\mathcal{M}}+\boldsymbol{w}_{\mathcal{M}}^{\perp}$と直交分解できるため，この 2 つのベクトルの内積$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$は$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}=\boldsymbol{w}_{\mathcal{M}}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{M}}+(\boldsymbol{w}_{\mathcal{M}}^{\perp})^{\mathrm{T}}\boldsymbol{x}_{\mathcal{M}}^{\perp}$と表すことができる．文献[8]の中では，この直交補空間成分を示す第 2 項が攻撃発生の要因としており，この成分に着目して本稿における提案手法を検討した．

4. 提案手法

提案手法では，敵対的サンプルを検知するために，データ多様体の直交補空間成分に着目する．すなわち，入力データがデータ多様体の接空間に対する直交補空間成分を持つか否かを判別することで，正常入力であるかそうでないかを判別する．ここではコサイン類似度を用いて，入力データの接空間に対する直交補空間成分を算出する．

検知・防御の立場では，学習データを取得することは可能であることを前提に，学習データ等をデータ多様体Mとして用いる．まず，Mを 1 次元高い空間に埋め込んだデータ多様体を$\mathcal{M}$とする．入力データ$\boldsymbol{x}$も同様に 1 次元高い空間に埋め込み$\boldsymbol{x}$とする．$\boldsymbol{x}$の最近点を$\mathcal{M}$上から選択し，最近点と同じラベルを持つ点を，$\boldsymbol{x}$を中心として近い順にk点選択することで近傍を作成する．この近傍を最近点で中心化することで局所近傍を得る．この局所近傍において主成分分析(PCA)を適用し，主成分と各主成分に対する寄与率を算出する．寄与率の累積和が閾値を超えたときの主成分の数を，局所次元mとし，m番目までの主成分軸$\boldsymbol{u}_i, i=1,\ldots,m$を接空間の基底として採用し，$U_{\mathcal{M}}=[\boldsymbol{u}_1,\ldots,\boldsymbol{u}_m]$とする．

最近点から$\boldsymbol{x}$へのベクトルを$\boldsymbol{v}$とする．$\boldsymbol{v}$の接空間への射影ベクトルを$\boldsymbol{v}_{\mathcal{M}}=U_{\mathcal{M}}U_{\mathcal{M}}^{T}\boldsymbol{v}$として，$\boldsymbol{v}$と$\boldsymbol{v}_{\mathcal{M}}$のコサイン類似度$\alpha$は

$$\alpha=\frac{\boldsymbol{v}\cdot\boldsymbol{v}_{\mathcal{M}}}{\|\boldsymbol{v}\|\|\boldsymbol{v}_{\mathcal{M}}\|}$$

と表すことができ，αが閾値を超えていれば$\boldsymbol{x}$がM上にあるため正常データ，そうでなければ$\boldsymbol{x}$のMに対する直交補空間成分が大きいため敵対的サンプルであると判定する．

5. 実験

正常画像と攻撃画像である敵対的サンプルを入力し，提案手法の性能評価を行った．

5.1 実験設定

本稿では縦横 28 ピクセルの手書き文字データセットの MNIST を拡張した QMNIST データセットを用いた[9]．正常入力には QMNIST の検証データの一部(10,000 点)，攻撃入力には，検証データのうち正しく分類された画像のみを用いて敵対的サンプルを生成し，さらに誤分類を引き起こした画像のみを利用した．敵対的サンプルの生成には CleverHans Library Ver.4.0.0[10]内の FGSM[11]を利用した．

ニューラルネットワークは 3 層の多層パーセプトロンであり、入力層は 784 ノード，中間層は活性化関数にシグモイド関数を用いた 200 ノードの全結合層，出力層は活性化関数にソフトマックス関数を用いた 10 ノードの全結合層で構成した．さらに QMNIST の訓練データ(60,000 点)を用いて学習し，予測精度は 98.03%であった．

QMNIST の訓練データ(60,000 点)に QMNIST の正常入力として使用していない検証データ(50,000 点)を加えたデータ集合(110,000 点)をデータ多様体とした．

5.2 実験結果

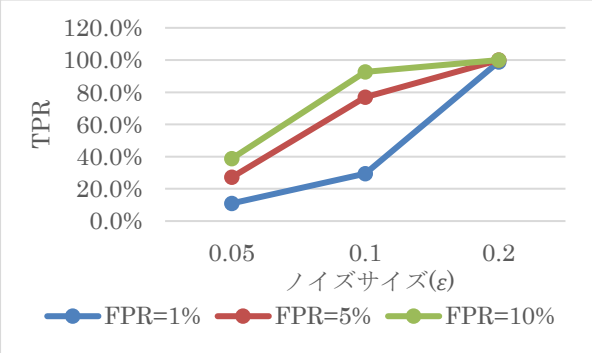

図.1 FGSM に対する提案手法の TPR

図.1 は，FGSM を利用して生成された攻撃サイズの違う攻撃画像と正常画像を入力した際の検知精度を表している．本稿では，偽陽性率(FPR)ごとの真陽性率(TPR)を用いて検知精度の評価を行った．この表からは，攻撃サイズが大きくなるにつれて検知精度が向上していることが分かる．

また図.2 は，$\varepsilon = 0.20$の攻撃画像を用いた際のコサイン類似度の分布を示したものであり，正常入力は青，攻撃入力は赤で表している．この図から，正常入力はデータ多様体の接空間とのコサイン類似度が 1 に近いこと，攻撃データは正常データより小さい値を持つ傾向があることから，提案手法が有効であることが確認できる．

6. 考察と今後の課題

本稿では，文献[8]で提案された敵対的サンプルの発生メカニズムに基づいた検知手法を提案し，その性能評価を行った．実験では，提案手法の有効性と，攻撃サイズが大きくなるにつれて検知精度が向上することが確認された．

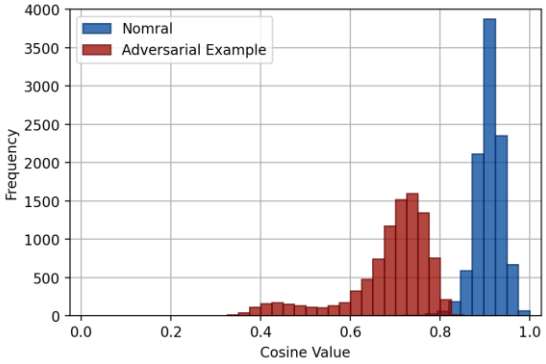

図.2 コサイン類似度のヒストグラム($\varepsilon = 0.20$)

しかし，攻撃サイズが小さいときに検知精度が低下した．この問題は，データ多様体に用いたデータ集合に攻撃の元データが存在していない場合に，データ集合内の最も近い点を元データとして採用したときに生じる誤差が原因である考えられる．

今後の課題として，攻撃サイズの小さな攻撃入力に対しての検知精度の向上が挙げられる．具体的に，データ多様体として用いるデータ集合のデータ密度を高くすることや，適切な接空間の推定方法により，改善が可能であると考えられる．また本稿では入力データがデータ多様体上に存在するかの指標としてコサイン類似度を利用したが，他の指標も検討すると同時に，他の攻撃手法やデータセットを用いた調査を続ける．また，最近点のラベルが攻撃サイズの小さい攻撃入力の正解ラベルと一致していることが多いため，これを利用した対策手法の検討も進める．

参考文献

[1] 産業技術総合研究所, “機械学習品質マネジメントガイドライン,” 第 2 版(revision 2.1.0), 2021.
[2] ENISA, “Cybersecurity Challenges in the Uptake of Artificial Intelligence in Autonomous Driving,” 2021.
[3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. “Intriguing properties of neural networks,” arXiv preprint arXiv:1312.6199, 2014.
[4] H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang, and A.K. Jain, “Adversarial attacks and defenses in images, graphs and text: A review”, International Journal of Automation and Computing, vol.17, no.2, pp.151–178 (2020).
[5] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, “A survey on adversarial attacks and defences,” CAAI Transactions on Intelligence Technology, vol.6, no.1, pp.25–45, 2021.
[6] N. Carlini, D. Wagner, “Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods,” arXiv preprint arXiv:1705.07263, 2017.
[7] K. Roth, Y. Kilcher, T. Hofmann, “The Odds are Odd: A Statistical Test for Detecting Adversarial Examples,” arXiv preprint arXiv:1902.04818, 2019.
[8] 田崎 元, 金子 勇次, 趙 晋輝, “埋め込み空間におけるデータ多様体構造に基づく敵対的サンプルの発生メカニズムに関する考察,” 信学技報, vol.121, no.192, PRMU2021-10, pp17-21, 2021.
[9] C. Yadav, L. Bottou, “Cold Case: the Lost MNIST Digits,” arXiv preprint arXiv:1905.10498, 2019.
[10] N. Papernot, F. Faghri, et al. “Technical report on the cleverhans v2.1.0 adversarial examples library,” arXiv preprint arXiv:1610.00768, 2018.
[11] I.J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” arXiv preprint arXiv:1412.6572, 2014.