F-036

データ多様体の埋め込み幾何学に基づく新しい敵対的サンプルによる攻撃手法

A New Attack to DNN by Adversarial Examples Based on Embedding Geometry of Data Manifolds

森田 匡博† 田崎 元† 趙 晋輝†
Masahiro Morita Hajime Tasaki Jinhui Chao

1. 序論

近年，深層学習を利用したサービスが増加しており，画像認識や音声認識，顔認証や自然言語処理など，幅広く展開されている．しかし，画像認識などでは，人間が知覚できないほど小さなノイズを加えることで誤分類を起こす敵対的サンプルの存在が報告されている[1]．この事例は産業技術総合研究所が出した機械学習品質マネジメントガイドラインで言及されている[2]．また，特定分野では EU の専門機関の一つである ENISA が発行した自動運転車のサイバーセキュリティの課題についての報告書にて，敵対的サンプルに対する対処法や注意事項が記載されるなど，その重要性が高まってきている[3].

なぜ敵対的サンプルが存在するのかについて長らく議論されていたが，確証ある理論に至らなかった．しかし，最近，学習データが持つデータ多様体の埋め込み構造を解析することで，敵対的サンプルはデータ多様体の接空間の直交補空間方向に存在することが明らかにされた[4].

そこで本研究では，上述の発生メカニズムに基づき，埋め込み空間におけるデータ多様体構造に着目した新たな敵対的サンプルの生成手法を提案する．提案手法では，多様体の直交補空間方向に，ノイズを加えることで攻撃画像を生成する．更に，多様体直交補空間成分が顕著な重みベクトル方向にノイズを加えることで攻撃画像を生成する手法を検討し，攻撃可能性について評価を行う．

2. 敵対的サンプルに関する先行研究

2.1 攻撃手法

2.1.1 Szegedy 法

2014 年に Szegedy らによって入力画像に微小なノイズを加えることで，ニューラルネットワークの誤分類を引き起こされることが指摘された[1]．敵対的サンプルは

$$\begin{aligned} &\text{Minimize} && c\|r\|_2 + \mathrm{loss}_f(x+r,l) \\ &\text{s.t.} && x+r \in [0,1]^m \end{aligned}$$

で定義された Box-constrained L-BFGS を解くことで生成される．ここでcは 0 以上の定数，rは摂動のサイズ，loss_fは損失関数，lはターゲットラベルである．様々な定数cに対して，ターゲットラベルlへの損失関数が最小に，また摂動も同時に最小になるような敵対的サンプルを上式によって探す.

2.1.2 FGSM

2014 年に Goodfellow らはニューラルネットワークの誤差関数の勾配情報を用いた Fast Gradient Sign Method(FGSM)を提案した[5]．ある入力画像xを元に生成された，敵対的サンプルx'は

$$x' = x + \epsilon \mathrm{sign}\big(\nabla_x J(\theta, x, y)\big)$$

で生成される．ここでθはモデルのパラメータ，$J(\theta,x,y)$は損失関数，ϵは摂動のサイズを制御するための定数である．FGSM は各ピクセルについて，損失関数の勾配を用いて，ターゲットラベルyに対する損失関数を最大化するための方向を 1 ステップで計算することで，最適化法である Szegedy 法と比較すると高速に生成することが可能となっている．

2.2 データの多様体構造

ニューラルネットワークの分類対象となる一般的な画像データは，画素数を次元とする高次元ベクトルで表現されるが，空間を充満しておらず，低次元の多様体構造を成す多様体仮説と呼ばれる性質を持つ.

2.3 敵対的サンプルの発生メカニズム

2021 年に田崎らはこれまで議論されてきた敵対的サンプルの発生メカニズムについて，データ多様体の空間構造に着目した理論を提唱した[4].

識別問題で扱われるデータはn次元の埋め込み空間 S に存在し，データ集合は低次元の多様体構造を持つ．これをd次元のデータ多様体Mとする．この多様体は局所的な近傍で線形近似することが可能であり，接空間というアフィン空間の集まりで表現することができる.

次に$n+1$次元の線形空間$\mathcal{S}$を考えると，データ多様体Mは$\mathcal{M} = \{\boldsymbol{x} = (\boldsymbol{x}^{\mathrm{T}}, 1)^{\mathrm{T}} \,|\, \boldsymbol{x} \in M\}$，重みは$\boldsymbol{w} = (\boldsymbol{w}^{\mathrm{T}}, \theta)^{\mathrm{T}}$となる．$\mathcal{M}$上のある点$\boldsymbol{x}$における接空間$T_{\boldsymbol{x}}\mathcal{M}$とそれに直交する空間である直交補空間$T_{\boldsymbol{x}}^{\perp}\mathcal{M}$への直交分解$T_{\boldsymbol{x}}\mathcal{S} = T_{\boldsymbol{x}}\mathcal{M} \oplus T_{\boldsymbol{x}}^{\perp}\mathcal{M}$が成立する．これにより，点$\boldsymbol{x}$は$\boldsymbol{x} = \boldsymbol{x}_{\mathcal{M}} + \boldsymbol{x}_{\mathcal{M}}^{\perp}$，重みも同様に$\boldsymbol{w} = \boldsymbol{w}_{\mathcal{M}} + \boldsymbol{w}_{\mathcal{M}}^{\perp}$と接空間と直交補空間の成分へそれぞれ分解することができる．よって，データ$\boldsymbol{x}$に対する重みとの内積は$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{w}_{\mathcal{M}}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{M}} + (\boldsymbol{w}_{\mathcal{M}}^{\perp})^{\mathrm{T}}\boldsymbol{x}_{\mathcal{M}}^{\perp}$と表せる．正常学習データでは$\boldsymbol{x}_{\mathcal{M}}^{\perp} = 0$であり，多様体方向の成分は，正常入力間の変形を表すため，人間に気づかれないと定義される敵対的サンプルにはほぼ含まれない．したがって，直交補空間へノイズが載った敵対的サンプルが入力された際は，第 2 項が十分大きな値になり，誤分類を引き起こす.

3. 提案手法

提案手法では，入力データと重みとの内積の第 2 項に着目し，内積を最大化する場合，つまり重みの直交補空間成分との内積が最大化されるようなデータ多様体の直交補空間成分を摂動として利用する.

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

まず MNIST の訓練データ集合をデータ多様体Mとし，それらをすべて 1 次元高い射影空間に埋め込んだデータ多様体を$\mathcal{M}$とする．攻撃元となるデータ$\boldsymbol{x} \in M$について埋め込み空間$\mathcal{S}$上でk近傍を作成し，$\boldsymbol{x}$で局所化を行う．次に主成分分析(PCA)を局所近傍で行い，ある閾値φの累積寄与率で算出された主成分または固有ベクトルを$\boldsymbol{u}_i$とし，接空間$T_{\boldsymbol{x}}\mathcal{M}$の直交基底を形成する．なお，$U_{\mathcal{M}} = [\boldsymbol{u}_1, \ldots \boldsymbol{u}_d]$であり，$d$は PCA での接空間の推定次元である．よって，摂動は以下の(1)式の第 2 項で求められる．

$$\begin{aligned} \boldsymbol{x}' &= \boldsymbol{x} \pm \varepsilon \frac{U_{\mathcal{M}}^{\perp}(U_{\mathcal{M}}^{\perp})^{\mathrm{T}} W \mathbf{1}_N}{\| U_{\mathcal{M}}^{\perp}(U_{\mathcal{M}}^{\perp})^{\mathrm{T}} W \mathbf{1}_N \|_2} \\ &= \boldsymbol{x} \pm \varepsilon \frac{W_{\mathcal{M}}^{\perp} \mathbf{1}_N}{\| W_{\mathcal{M}}^{\perp} \mathbf{1}_N \|_2} \end{aligned} \tag{1}$$

$W_{\mathcal{M}}^{\perp}$の列ベクトルは多様体の直交補空間方向へ射影された重みの直交補空間ベクトル全体であり，$W_{\mathcal{M}}^{\perp} = W - W_{\mathcal{M}} = W - U_{\mathcal{M}}(U_{\mathcal{M}})^{\mathrm{T}} W$と変形できる．$\mathbf{1}_N$は全要素が 1 の$N$次元列ベクトルである．ここで，$N$は中間層のノード数である．なお，摂動である第 2 項は埋め込み空間$\mathcal{S}$上のベクトルなので，先にアフィン空間 Sに戻す必要がある．

攻撃元画像$\boldsymbol{x}$に接空間$T_{\boldsymbol{x}}\mathcal{M}$に直交する直交補空間$T_{\boldsymbol{x}}^{\perp}\mathcal{M}$の直交基底集合$U_{\mathcal{M}}^{\perp}$の成分を摂動としてアフィン空間 S上で加算すれば敵対的サンプル$\boldsymbol{x}'$の生成が可能である．さらに，中間層のi番目のニューロンの重みベクトル$\boldsymbol{w}_i$との内積を取り，最大値を取るような直交補空間の直交基底を摂動として選べばより効果的な攻撃が可能である．これは，重みベクトルの直交補空間方向を画像に加えることと同義である．また，(1)式で敵対的サンプル$\boldsymbol{x}'$を生成した後に，画像のピクセル値の範囲を 0 以上 1 以下にするクリップ処理を行う．その際，起こりうる摂動の変化を抑えるため，負のピクセル値の数がより少なくなるような符号を決める．

4. 実験

データ多様体$\mathcal{M}$の直交補空間$T_{\boldsymbol{x}}^{\perp}\mathcal{M}$を利用することにより，敵対的サンプルの発生メカニズムに基づいた提案手法の有効性を検証した．

4.1 実験設定

攻撃の元となる画像は MNIST 内の訓練データ集合の中から 1 万枚を利用し，攻撃の比較対象として FGSM を利用するにあたり，CleverHans Library Ver.3.10[6]を利用した．また，対象のニューラルネットワークには多層パーセプトロンを利用し，入力層は 784 ノード，中間層はシグモイド関数を活性化関数とする 200 ノードの全結合層を 1 層とし，出力層はソフトマックス関数を活性化関数とする 10 ノードの全結合層で構成した．なお，分類精度は 98%である．

攻撃の生成においては，1 つの近傍に属する数を 150 個，接空間の推定における PCA の累積寄与率$\varphi = 0.99$とする．

図 1（左）元画像，（右）生成画像

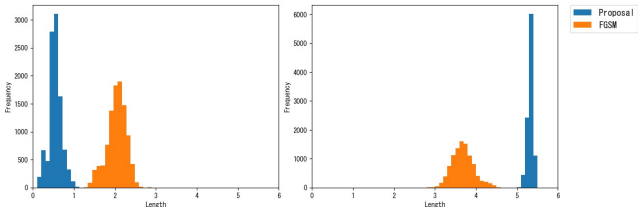

図 2 (a) $\|\boldsymbol{r}_{\mathcal{M}}\|_2$の分布，(b) $\|\boldsymbol{r}_{\mathcal{M}}^{\perp}\|_2$の分布

4.2 実験結果

4.2.1 攻撃画像の生成

提案手法では，$\varepsilon = 0.2$に設定した場合の攻撃成功率は 89%であり，生成された敵対的サンプルの画像を図 1 に示す．ここでは，生成画像は 6 を 2 と誤分類している．提案手法で作成された攻撃画像には，図 1 のような字体の変形ではない背景雑音が含まれる傾向が確認できた．

4.2.2 既存手法との比較

FGSM の生成画像と，提案手法の生成画像それぞれに対して，埋め込み空間$\mathcal{S}$上で摂動$\boldsymbol{r} = \tilde{\boldsymbol{x}} - \boldsymbol{x}$を求め，接空間成分$\boldsymbol{r}_{\mathcal{M}}$の長さ$\|\boldsymbol{r}_{\mathcal{M}}\|_2$と，直交補空間成分$\boldsymbol{r}_{\mathcal{M}}^{\perp}$の長さ$\|\boldsymbol{r}_{\mathcal{M}}^{\perp}\|_2$をそれぞれ比較したヒストグラムを図 2 に示す．なお，FGSM の攻撃成功率は 99%である．

図 2 の(a)(b)より，FGSM は提案手法に比べて，直交補空間成分が小さいが，多様体成分が大きいことが分かる．FGSM には，誤分類を起こすために多様体成分$\boldsymbol{r}_{\mathcal{M}}$が含まれており，厳密には敵対的サンプルとは異なる攻撃も含まれていると考えられる．また，重みの接空間成分$\boldsymbol{w}_{\mathcal{M}}$は，学習の際のランダム成長しか期待できない直交補空間成分$\boldsymbol{w}_{\mathcal{M}}^{\perp}$よりも値が大きいので，内積計算の際における摂動の接空間成分$\boldsymbol{r}_{\mathcal{M}}$の分類への影響も大きくなると思われる．

5. 結論と今後の課題

先行研究で提唱された敵対的サンプルの発生メカニズムを用いた攻撃画像の生成法を提案して，その有効性を確認した．今後の課題として他の攻撃との比較評価を行い，さらに，敵対攻撃の可視性に関する調査およびそれを含む評価方式に関しての検討が挙げられる．

参考文献

[1] 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).
[2] 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”, (2021).
[3] G. Dede, R. Hamon, R. Naydenov, H. Junklewitz, A. Malatras, I. Sanchez, “Cybersecurity Challenges in the Uptake of Artificial Intelligence in Autonomous Driving”, pp.32-35 (2021).
[4] 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, pp. 17-21 (2021).
[5] I.J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples”, arXiv preprint arXiv:1412.6572, (2014).
[6] N. Papernot, F. Faghri, N. Carlini, I. Goodfellow, R. Feinman, A. Kurakin, C. Xie, Y. Sharma, T. Brown, A. Roy, A. Matyasko, V. Behzadan, K. Hambardzumyan, Z. Zhang, Y.-L. Juang, Z. Li, R. Sheatsley, A. Garg, J. Uesato, W. Gierke, Y. Dong, D. Berthelot, P. Hendricks, J. Rauber, and R. Long, “Technical report on the cleverhans v2.1.0 adversarial examples library,” arXiv preprint arXiv:1610.00768, (2018).