H-033

AdvGAN を用いた敵対的サンプル攻撃に対するデータ多様体に基づく解析と防御手法

Analysis and Defense Based on Data Manifold for Adversarial Attack with AdvGAN

王 翰† 田崎 元† 趙 晋輝†
Wang Han Tasaki Hajime Jinhui Chao

1. はじめに

近年，さまざまな領域で用いられる深層学習は，小さな摂動を加えた不正な入力である敵対的サンプルによって，簡単に誤分類が引き起こされることが知られている．この敵対的サンプルを生成する攻撃手法は数多く研究されており，従来のニューラルネットワークの損失関数に対する勾配を用いた生成手法や摂動の最適化による生成手法に加え，敵対的生成ネットワーク（GAN）を用いた攻撃手法も提案されている．本研究では，GAN を用いた攻撃手法である AdvGAN を対象に，著者らが提案してきたデータ多様体に対する埋め込み幾何学の理論に基づく解析を適用し，防御手法を提案する．

2. 関連研究

2.1 GAN と AdvGAN

GAN（Generative Adversarial Network）は Goodfellow らによって提案され，生成ネットワーク（Generator）と識別ネットワーク（Discriminator）の２つのネットワークを互いに競わせながら学習を行う[1].

AdvGAN（Adversarial GAN）は Xiao らが提案した敵対的サンプルを生成する GAN である[2]．AdvGAN は生成ネットワークG，識別ネットワークD，攻撃対象のニューラルネットワークfの３つのネットワークを持つ．図 1 に AdvGAN の構成を示す．AdvGAN は，LSGAN（Least Square GAN）の損失関数を用いた学習が不安定であるため，Xiong らは，LSGAN より学習安定性が優れる WGAN-GP[3]の損失関数を AdvGAN に応用した AdvGAN の改良版[4]を提案した．

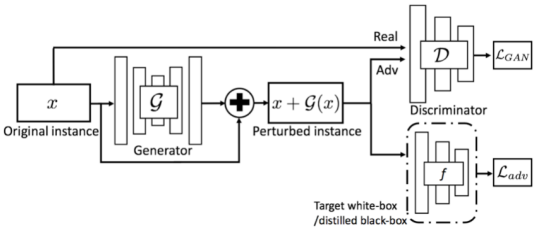

図 1 AdvGAN の構成[2]

2.2 敵対的サンプルの発生メカニズム

敵対的サンプルの発生メカニズムは文献[5]及び文献[6]で提案されている．画像データのデータ集合は低次元多様体構造を持つことを仮定し，データ及び重みはデータの変形成分である多様体の接空間成分と，それ以外の直交補空間成分に分解できることを明らかにした．また，正常データのみから構成される学習データは，多様体の直交補空間成分は持たないことから，重みの直交補空間方向の成分は学習することができず，適切に制御することは不可能である．その結果，直交補空間成分を持った敵対的サンプルに対して，ニューラルネットワークは誤分類を起こす.

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

3. 提案手法

文献[7]及び文献[8]に基づく敵対的サンプルの防御手法と解析手法を提案する．

防御手法　学習データは点群$\boldsymbol{X} = \{\boldsymbol{x}_i | \boldsymbol{x}_i \in M, i = 1, \dots, N\}$でデータ多様体$\boldsymbol{M}$を構成し，射影空間$\mathcal{S}$に写像された点群$\mathbf{X} = \{\mathbf{x}_i = \left(\boldsymbol{x}_i^{\mathrm{T}}, 1\right)^T | i = 1, \dots, N\}$は射影空間$\mathcal{S}$におけるデータ多様体$\boldsymbol{\mathcal{M}}$を構成する．同様にして，射影空間$\mathcal{S}$に埋め込まれた入力データ$\boldsymbol{x}$を$\mathbf{x} \in \mathbb{R}^{n+1}$とする．入力データ$\mathbf{x}$の最近点$\mathbf{x}_0$をデータ多様体$\boldsymbol{\mathcal{M}}$から探索し，最近点$\mathbf{x}_0$と同じラベルを持つ点のうち，$\mathbf{x}_0$から近い順に$k$点を取得して，近傍を作成する．さらに主成分分析（PCA）を適用し，降順に並ぶ主成分$\sigma_1, \dots, \sigma_k$と対応する主成分ベクトル$\boldsymbol{u}_1, \dots, \boldsymbol{u}_k$を算出する．累積寄与率$\varphi_m = (\Sigma_{i=1}^{m} \sigma_i) / \left(\Sigma_{i=1}^{k} \sigma_i\right)$を計算し，閾値λに対して$\varphi_m \geq \lambda$を満たす主成分の数$m$を接空間の次元として推定し，$m$番目までの主成分ベクトルを接空間の基底$\boldsymbol{U}_{\mathcal{M}} = [\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_m]$とする．入力データ$\boldsymbol{x}$の接空間の方向成分$\mathbf{x}_{\mathcal{M}} = \boldsymbol{U}_{\mathcal{M}} \boldsymbol{U}_{\mathcal{M}}^{T} (\mathbf{x} - \mathbf{x}_0) + \mathbf{x}_0$を抽出する．接空間の方向成分$\mathbf{x}_{\mathcal{M}} \in \mathcal{S}$に対して，$\boldsymbol{x}' = (\mathrm{x}_1 / \mathrm{x}_{n+1}, \dots, \mathrm{x}_n / \mathrm{x}_{n+1})$を計算し，アフィン空間$S$における座標表現を得る．得られた$\boldsymbol{x}'$に対して分類を行うことで，敵対的サンプルによる誤分類を防ぐ．

解析手法　本解析手法は，生成された敵対的サンプルの摂動を接空間方向及び直交補空間方向の成分に分解し，各方向の摂動と識別境界の距離を調べることで，誤分類の起こしやすさを解析する手法である．防御手法と同様の操作を行うが，本解析は摂動に対する解析であり，近傍の推定精度による影響を取り除くため，最近点$\boldsymbol{x}_0$には敵対的サンプルの元画像を利用して摂動を抽出する．

まず，敵対的サンプル$\tilde{\boldsymbol{x}}$の接空間成分$\tilde{\boldsymbol{x}}_M$と直交補空間成分$\tilde{\boldsymbol{x}}_M^{\perp}$を抽出する．摂動の接空間成分$\boldsymbol{v}_M = \tilde{\boldsymbol{x}}_M - \boldsymbol{x}_0$と直交補空間成分$\boldsymbol{v}_M^{\perp} = \tilde{\boldsymbol{x}}_M^{\perp} - \boldsymbol{x}_0$の単位ベクトル$\boldsymbol{v'}_M$と$(\boldsymbol{v}_M^{\perp})'$を計算する．接空間方向の識別境界までの距離を測定するには，初期移動幅をϵとして，元画像$\boldsymbol{x}_0$に$\epsilon \boldsymbol{v'}_M$を加えていき，誤分類を引き起こすまで繰り返す．誤分類を起こしたとき，識別境界を越えたと判断して，移動幅ϵを$-\epsilon/2$に更新し，逆方向に小さい移動幅で同様の操作を行う．この移動幅$|\epsilon|$が閾値γを下回るまで識別境界の探索を繰り返し，収束した識別境界上の画像と元画像$\boldsymbol{x}_0$の距離を識別境界までの距離とする．直交補空間方向の識別境界を測定するには$\boldsymbol{v'}_M$を$(\boldsymbol{v}_M^{\perp})'$に置き換え，同様に求めることができる．

4. 実験と結果

4.1 実験設定

本実験では，MNIST データセットを用いて畳み込みニューラルネットワーク（CNN）と多層パーセプトロン（MLP）

に対する実験を行った．CNN はカーネルサイズが 3，ストライドが 1 の 4 層の畳み込み層と 1024 ノード，200 ノード，10 ノードを持つ 3 層の全結合層で構成される．MLP は 784 ノードを持つ入力層，512 ノードを持つ中間層と Softmax 関数を用いた 10 ノードを持つ出力層から構成される．両モデルの全結合層は ReLU を活性化関数とし，CNN では確率 $p = 0.2$，MLP では確率$p = 0.5$の Dropout を用いた．MNIST のテストデータに対するモデルの分類精度は 99.8%（CNN）と 99.6%（MLP）であった．

CNN と MLP を攻撃対象のニューラルネットワークfとし，AdvGAN の改良版（以下は AdvGAN と略す）を用いて，MNIST の 10,000 枚のテストデータに対する敵対的サンプルを生成した．生成された敵対的サンプルのうち，誤分類に成功した画像は，CNN では 9,585 枚，MLP では 9,615 枚であった．元画像と生成された画像を図 2 に示す．CNN に対する敵対的サンプルは，元の数字の外見を別の数字に変形させた画像となる傾向を持つことがわかった．一方で，MLP に対する敵対的サンプルは，数字の変形がはっきりとは現れず，ノイズに近い摂動が生成される．

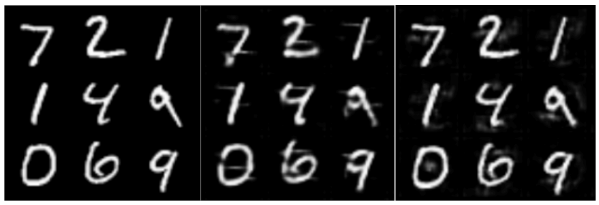

図 2 左：正常画像，中：敵対的サンプル（CNN），右：敵対的サンプル（MLP）

4.2 防御手法の検証

AdvGAN で生成された敵対的サンプルに対して，防御手法を適用し有効性を検証した．本実験では，累積寄与率λを 90%，近傍内の点数kを 100 点に設定し，評価指標には，

$$\text{Accuracy} = (\text{正しく分類できた数})/(\text{攻撃画像の総数})$$

を防御成功率として採用した．その結果，CNN に対する防御成功率は 93.7%，MLP に対する防御成功率は 93.8%で，いずれも高い精度で誤分類を防ぐことができた．

また，防御手法により抽出された$\boldsymbol{x}_M$の画像を図 3 に示す．図 3 のいずれの画像も，ノイズのように付加された摂動は除去することができており，より自然な画像を得られた．これらの結果より，本手法は AdvGAN による敵対的サンプルに対する防御手法として有効であることがわかった．

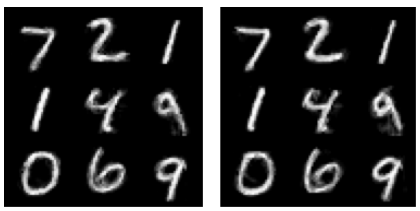

図 3 左：防御画像（CNN），右：防御画像（MLP）

4.3 敵対的サンプルの成分と攻撃可能性

次に，AdvGAN で生成される敵対的サンプルに含まれる摂動の方向と識別境界に関する解析を行った．成分ごとの誤分類の起こしやすさ調べるため，解析手法を用いて，摂動を接空間方向成分と直交補空間方向成分に分解し，各方向の識別境界への距離の分布を，図 4 の頻度分布に示す．図 4 は，縦軸は接空間方向の識別境界への距離を示しており，横軸は直交補空間方向の識別境界への距離を示している．本実験では，移動幅の初期値ϵを0.1，更新の終了条件γを$1e-7$に設定し，解析手法を適用した．その結果，CNN は横軸方向に比べ，縦軸方向に広く分布しており，識別境界までの距離が長い傾向にあることがわかった．MLP は CNN よりも分布が集中しているものの，CNN と同様，縦軸方向に広く分布する傾向が見られた．この結果から，接空間方向に識別境界を越えるには，一定量の摂動が必要であるが，直交補空間方向には，それよりも少ない摂動で識別境界を越えることから，AdvGAN による敵対的サンプルから直交補空間方向の識別境界までの距離は短いことがわかる．したがって，防御手法は，誤分類への影響が大きい直交補空間方向の成分を取り除くことができるため，高い成功率で誤分類を防ぐことができたと考えられる．

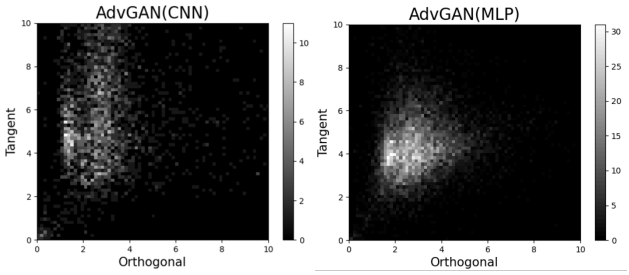

図 4 識別境界までの距離の頻度分布

5. まとめ

本稿では，AdvGAN による敵対的サンプルに対して，田崎らが提案した防御手法[8]を適用し，AdvGAN による敵対的サンプルに対して高い成功率で防御できることを実証した．加えて，データ多様体の埋め込み幾何学の観点から，データ解析を行い，誤分類を起こす要因である直交補空間成分を除去することに優れることから，高い精度で防御できることを明らかにした．今後の課題として，大規模なデータセットで AdvGAN による敵対的サンプルに対する解析と防御の実験が挙げられる．

謝辞

本稿の執筆にあたり，有益な助言，議論及び協力を頂いた中央大学理工学研究科 久重広樹氏に感謝したい．

参考文献

[1] I. Goodfellow et al. "Generative adversarial nets",arXiv:1406.2661, 2014.
[2] C.W. Xiao et al. "Generating Adversarial Examples with Adversarial Networks",IJCAI,2018.
[3] I. Gulrajani et al. "Improved Training of Wasserstein GANs", arXiv:1704.00028,2017
[4] Z.H. Xiong et al. "Using Wasserstein GAN to Generate Adversarial Examples",http://hdl.handle.net/2142/100047,2018.
[5] 田崎 他．"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",PRMU2021.
[6] H.Tasaki et al. "Curse of co-Dimensionality: Explaining Adversarial Examples by Embedding Geometry of Data Manifold". ICPR 2022.
[7] 久重 他. "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",FIT2022.
[8] 田崎 他. "敵対的サンプル攻撃に対するデータ多様体の埋め込み幾何学に基づく防御手法",FIT2022.