CH-010

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
Defense for DNN against Adversarial Example Attacks based on Embedding Geometry of Data Manifolds

田崎 元† 趙 晋輝†
Hajime Tasaki Jinhui Chao

1. 序論

ニューラルネットワーク（以降，NN と記述する）は，近年の著しい発展を遂げたAI技術，とりわけ深層学習を支える中核技術である．その一方で，NN を用いた分類器に対して，故意に誤った予測をさせる不正なデータである敵対的サンプルが発見されている[1]．敵対的サンプルは，摂動と呼ばれる微小なノイズをデータに付加することで生成され，観測者が正規データとの違いを認知することは困難であることから，NN の脆弱性として問題視されている[2][3]．図 1 が CIFAR10 データセットを用いて，実際に生成した敵対的サンプルの一例である．

この敵対的サンプルによる影響は，AI 技術を利用するシステムやサービスの信頼性に直結することから，産業技術総合研究所から機械学習品質マネジメントガイドライン[4]が発行されている．また，特定分野では欧州連合の一機関である ENISA から自動運転における AI 技術に対するセキュリティレポート[5]が発行されており，いずれも敵対的サンプルに対する注意事項や対策手法が盛り込まれている．

敵対的サンプルが発見された当時，ミステリアスな現象と呼ばれ，その発生原理について議論されてきた．その中で有力な説として，Goodfellow らが指摘した高次元の内積計算[5]や，汎化や学習の誤りが原因と指摘された．しかしながら，それらの仮説では，この現象を十分に説明し切ることができなかった上に，それらに基づく有効な対策手法は打ち出されてこなかった．そこで，筆者らは学習データの多様体構造に着目した新たな発生メカニズムを発表している[9][10]．この文献[10]において，従来は解明できなかった敵対的サンプルの特性や高次元空間における敵対的サンプルの存在領域などに対して生じていた矛盾が解消できることを示した．

本研究は，この発生メカニズムに基づき，学習データの多様体成分を抽出し，入力データから攻撃の要因となる多様体の直交成分を除去することにより，敵対的サンプルによる誤分類を防ぐ新たな防御手法を提案する．

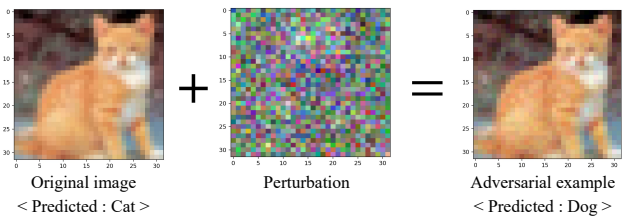

図 1 敵対的サンプルの一例（[11]より）

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

2. 敵対的サンプルに関わる先行研究

2.1 攻撃手法

敵対的サンプルは，観測者には知覚困難なほど微小な摂動を付加することで生成される．これを定式化すると，敵対的サンプル$\tilde{\boldsymbol{x}}$は，正常画像を$\boldsymbol{x}$と摂動$\boldsymbol{r}$を用いて，

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{r} \tag{1}$$

と表せる．この摂動$\boldsymbol{r}$の生成方法には，さまざまな手法があるが，最適化に基づく手法と，NN で用いられる損失関数の勾配を用いる手法に大きく分けられる．本稿では，それぞれの最も基本となる手法として，Szegedy 法と FGSM 法を述べる．

Szegedy 法

Szegedy らは，敵対的サンプルの摂動生成を最適化問題として定式化し，その近似解を導出することで敵対的サンプルを求める手法を提案した．この手法では，関数fを$f: \mathbb{R}^n \to \{1, \dots, k\}$とする判別関数とし，損失関数を$\mathrm{loss}_f: \mathbb{R}^n \times \{1, \dots, k\} \to \mathbb{R}$と表すとき，誤ったクラスへ識別させることを条件とした敵対的サンプルの摂動に対する最小化問題として定式化している．

$$\begin{aligned} &\text{Minimize} && \|\boldsymbol{r}\|_2 \\ &\text{subject to} && f(\boldsymbol{x}+\boldsymbol{r}) = l \; and \; \boldsymbol{x}+\boldsymbol{r} \in [0.0, 1.0]^n \end{aligned}$$

しかし，一般に，摂動$\boldsymbol{r}$を最適解として一意に定めることは困難であるため，最適化の目的関数に NN の損失関数を導入して，

$$\begin{aligned} &\text{Minimize} && c\|\boldsymbol{r}\|_2 + \mathrm{loss}(\boldsymbol{x}+\boldsymbol{r}, l) \\ &\text{subject to} && \boldsymbol{x}+\boldsymbol{r} \in [0.0, 1.0]^n \end{aligned}$$

という最適化問題に対して LBFGS により，反復的に求めることで近似解を得る．

FGSM 法：Fast Gradient Sign Method

FGSM 法は，NN の分類器における損失関数の勾配情報を用いて，摂動を生成する手法である．この手法では，損失関数 $J(\boldsymbol{w}; \boldsymbol{x}, r)$をデータ$x$のラベル$y$に対する NN の損失関数として，摂動$r$を

$$\boldsymbol{r} = \varepsilon \, \mathrm{sign}\big(\nabla_{\boldsymbol{x}} J(\boldsymbol{w}; \boldsymbol{x}, y)\big) \tag{2}$$

と定義することにより，敵対的サンプル$\tilde{\boldsymbol{x}}$を

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \varepsilon \, \mathrm{sign}\big(\nabla_{\boldsymbol{x}} J(\boldsymbol{w}; \boldsymbol{x}, y)\big) \tag{3}$$

で求められると発表した．この手法は，NN のある重み$\boldsymbol{w}$におけるデータxに対する勾配情報が得られれば，1 ステップで敵対的サンプルを生成できる．Szegedy 法は，摂動の最小化により，知覚困難な敵対的サンプルを生成できるという観点で強い攻撃であるのに対して，FGSM 法で生成される画像は粗いものの，高速に敵対的サンプルを生成する手法と言われている．

2.2 防御手法

分類器の安全性を向上させるため，敵対的サンプルに対する対策手法が提案されている．対策手法には，摂動が付加されても正規の入力と同一の予測を行えるようにする防御手法と，入力に対してあらかじめ敵対的サンプルであるかを判定する検知手法の２つに大別される．後者では，敵対的サンプルと判定されれば識別を行わないなどのシナリオが考えられるが，本稿では，入力画像に対して正規の識別を実現することを目標とし，防御手法を議論の対象とする．防御手法は幅広く研究が行われており，提案された手法も多くあるが，その中でも防御の中でも広く知られる手法について紹介する．

まず，敵対的学習（Adversarial Training）が提案されている．この手法は，識別タスクを行うモデルの学習に，ある定められた損失関数の下，正しくラベル付けされた敵対的サンプルを訓練データに与えることで，誤分類を抑制する手法である．初期のコンセプトは Szegedy らによって発表され，以降も改善手法として FGSM 法を用いた学習方式[6]や，Projective Gradient Descent（PGD）を利用した Madry らの学習方式[7]が提案されている．これらの敵対的学習は数々の改良が重ねられ，いくつもの派生した学習方式が提案されている．しかしながら，敵対的学習は敵対的サンプルに対する学習を行うため，ノイズを含むサンプルから識別境界を学習する．したがって，正常データのみから学習された識別境界からノイズ方向に広がったものとなるため，正常入力に対する識別精度の低下を起こすことが課題と指摘されている[12]．また，筆者らの敵対的サンプルの発生メカニズムに基づけば，攻撃が生成される空間は，データ多様体外の空間であり，多様体の余次元は非常に高いことから，敵対的サンプルによる誤分類を防ぐために網羅的な学習をすることは困難である．さらには，学習ベースの防御手法では，未知な攻撃を対策することは，分類精度の劣化を許容しても，困難であると考えられる．

次に，Papernot らによって，Defensive distillation が提案されており，計算リソースに乏しい端末で深層学習を動作させることを目標に，ネットワークアーキテクチャを小さくするための蒸留という学習方式を応用した防御手法である[8]．この手法は，同じネットワーク構造の NN を 2 つ利用し，片方の NN の出力でロジットを出力するよう構成する．この出力されるロジットを，もう一方の NN の学習データに対する正解ラベルとして利用することで，分類精度の向上が可能とし，これにより，敵対的サンプルに対する耐性を大きく向上させると報告されている．

3. 提案手法

3.1 データ多様体と多様体仮説

深層学習が分類の対象とする高次元データは，一般に高次元空間に埋め込まれた非線形な低次元の多様体上に集中して存在する．この性質は“多様体仮説”と呼ばれている[12]．これはデータ点が，3 次元空間で形成する曲線や曲面のように，必ずしも空間を充満して分布するのではなく，低次元の限られた部分空間に存在することを意味している．これに基づき，低次元に分布するデータ集合に対する幾何学的構造の解析や特徴抽出を行うことで，そのデータ集合の持つ低次元の特徴表現を獲得するデータ分析手法が知られている．特に，画像を対象にした研究は古くから行われており，多様体構造を持つデータ集合として扱い，幾何学的解析のアプローチがとられている[15].

本研究においても，多様体仮説のもとで画像データセットを対象に，可微分多様体[16]からの標本として，その標本データの多様体をデータ多様体と呼ぶこととする．さらに，データセットは十分なサンプル密度をもち，局所的な連続性も担保されることを仮定し，データ多様体の接空間など，データセットに対する幾何学的解析を導入する．

3.2 敵対的サンプルの発生メカニズム

敵対的サンプルにおける摂動は，知覚困難なほどに小さなものでありながら，誤分類を誘発させる．この現象に関して，筆者らは NN の学習データによるデータ多様体に対して，敵対的サンプルは異なった特徴的な部分空間に存在することを文献[10]で発表している．本稿では，この発生メカニズムに基づく防御手法の提案であるため，その詳細について述べる．

3.1 節で述べたとおり，データ点$\boldsymbol{x}$の集合からなるデータ多様体Mはユークリッド空間$\mathbb{R}^n$であるデータ空間Sに埋め込まれたm次元部分多様体とする．

NN の各ノードにおける重みベクトル$\boldsymbol{w}$とバイアスθとして，入力データ$\boldsymbol{x}$に対する出力$\boldsymbol{y}$は，関数fを活性化関数として，$f(\boldsymbol{w}^\top \boldsymbol{x}+\theta)$である．ここで，データ空間$S$における点$x \in M$に対して，射影空間への変換を$\boldsymbol{x}=(\boldsymbol{x}^\top,1)^\top$とするとき，射影空間$\mathcal{S}:=\mathbb{R}^{n+1}$におけるテータ多様体$M$は，$\mathcal{M}=\{\boldsymbol{x}=(\boldsymbol{x}^\top,1)^\top|\ \boldsymbol{x}\in M\}$と表せて，同様にして重み$w$は$\boldsymbol{w}=(\boldsymbol{w}^\top,\theta)^\top$と表せる．これらを用いて，NN のノードにおける内積計算は$\boldsymbol{y}=f(\boldsymbol{w}^\top \boldsymbol{x})$となる．

データ多様体$\mathcal{M}$上の点$\boldsymbol{p}$に対する接空間$T_p\mathcal{S}$は，接空間$T_p\mathcal{M}$と，その直交補空間$T_p\mathcal{M}^\perp$を用いて，

$$T_p\mathcal{S}=T_p\mathcal{M}\oplus T_p\mathcal{M}^\perp \tag{3}$$

と直交分解することができる．さらには，点pの近傍におけるデータ点$\boldsymbol{x}$と重みベクトル$\boldsymbol{w}$は，同様にして，

$$\boldsymbol{x}=\boldsymbol{x}_{\mathcal{M}}+\boldsymbol{x}_{\mathcal{M}}^{\perp} \tag{4}$$

$$\boldsymbol{w}=\boldsymbol{w}_{\mathcal{M}}+\boldsymbol{w}_{\mathcal{M}}^{\perp} \tag{5}$$

として，多様体$\mathcal{M}$の接空間方向の成分と，その直交補空間方向の成分に分解可能であることを示した．この分解に基づけば，摂動も$\boldsymbol{r}=\boldsymbol{r}_{\mathcal{M}}+\boldsymbol{r}_{\mathcal{M}}^{\perp}$と表せて，敵対的サンプルに対する NN の内積計算は，

$$\boldsymbol{w}^\intercal\tilde{\boldsymbol{x}}=\boldsymbol{w}_{\mathcal{M}}^\intercal\boldsymbol{x}_{\mathcal{M}}+\boldsymbol{w}_{\mathcal{M}}^\intercal\boldsymbol{r}_{\mathcal{M}}+\left(\boldsymbol{w}_{\mathcal{M}}^{\perp}\right)^\top\boldsymbol{r}_{\mathcal{M}}^{\perp} \tag{6}$$

となる．この式は，一般に，データ多様体$\mathcal{M}$の接空間の方向は，データの変化を表す方向であり，観測者に識別されにくい敵対的サンプルを生成するためには，$r\approx r_{\mathcal{M}}^{\perp}$と考えられ，入力データ$\boldsymbol{x}$はオリジナルの正規データとすれば，データ多様体$\mathcal{M}$上に存在するため，$\boldsymbol{x}_{\mathcal{M}}^{\perp}\approx\boldsymbol{0}$となることに基づいている．

3.3 データ多様体に基づく敵対的サンプルの防御手法

本稿では，敵対的サンプルの発生メカニズムに基づき，データ多様体の空間構造に着目した新たな防御手法を提案する．

3.2 節で述べたとおり，データ多様体$\mathcal{M}$は直交分解により，接空間$T_p\mathcal{M}$とその直交補空間$T_p\mathcal{M}^\perp$に分解可能である．データ多様体$\mathcal{M}$の分布は，文字画像ならば字体の変形，表情画像ならば表情変化などのようなデータの変形，すなわ

ち，分類タスクにおいて注目すべき潜在的な特徴を意味している．したがって，敵対的サンプルが人間に知覚されにくいことを前提とすれば，接空間方向に摂動を付加するのではなく，直交補空間の方向に摂動が付加することが敵対的サンプルの生成に寄与すると考えられる．

そのため，提案手法では式(6)における第３項を取り除くことが有効であると考え，データ多様体の空間構造に注目し，入力データの直交補空間方向の成分を取り除くことで，敵対的サンプルの元となる正規の入力に対する分類結果を得ることを目的とする．

提案手法におけるアルゴリズムの流れを説明するにあたり，データ多様体Mを構成する点群$\boldsymbol{X} = \{\boldsymbol{x}_i | \, \boldsymbol{x}_i \in M, i = 1, \ldots, N\}$を射影空間$\mathcal{S}$へ埋め込み，$\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}_i = (\boldsymbol{x}_i^\top, 1)^\top | \, i = 1, \ldots, N\}$を事前準備として算出しておく．本稿では，データ多様体Mは NN の学習データで構成されることを想定する．以下に，分類対象とする入力データ$\boldsymbol{x}$に対する処理の流れを示す．

1. 入力データ$\boldsymbol{x} \in \mathbb{R}^n$を射影空間$\mathcal{S}$に埋め込み，$\boldsymbol{x} \in \mathbb{R}^{n+1}$を得る．
2. 入力データ$\boldsymbol{x}$からの距離が最短となる点を点群$\mathcal{X}$から探索し，最短距離の点から距離が近い順にk点を取得し，$\boldsymbol{x}$の近傍とする．
3. 得られた近傍において，特異値分解を適用し，特異値の累積寄与率φが 99%を超える特異値の個数mをデータ多様体の接空間の次元$\dim(T_x\mathcal{M}) = m$とし，特異値が大きい順に対応する特異ベクトル$\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m$による行列$\boldsymbol{U}_{\mathcal{M}} = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_m]$を求める．
4. 得られた特異ベクトルを接空間$T_{\boldsymbol{x}}M$の基底として，接空間に直交射影を行い，入力データ$\boldsymbol{x}$における接空間の方向成分$\boldsymbol{x}_{\mathcal{M}}$を抽出する．この$\boldsymbol{x}_{\mathcal{M}}$は，次式で求めることができる．
$$\boldsymbol{x}_{\mathcal{M}} = \boldsymbol{U}_{\mathcal{M}} \boldsymbol{U}_{\mathcal{M}}^\top \boldsymbol{x}$$
5. 得られた$\boldsymbol{x} \in \boldsymbol{\mathcal{S}}$の接空間方向の成分に対して，$\boldsymbol{x}_M = (\boldsymbol{x}_1 / \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_n / \boldsymbol{x}_{n+1})$を計算することにより，アフィン空間$S$における座標表現を得る．この入力データ$\boldsymbol{x}$の接方向成分$\boldsymbol{x}_M$を計算した上で，NN による分類を行うことにより，仮に入力データが摂動を含む敵対的サンプルであっても，正規の分類結果を得られることができる．

以上の手順に示すように，提案手法は NN の学習データさえ利用できればよく，従来手法のように NN の再学習は必要としないうえに，NN の重みや構成などのパラメータを利用できない場合においても，適用可能な手法である．さらには，正規の入力データ$\boldsymbol{x}$は$\boldsymbol{x} \approx \boldsymbol{x}_M$であると仮定すれば，提案手法を適用することによる正規入力に対する分類精度の大幅な劣化を起こさずに，防御手法の導入が期待できる．

4. 実験

4.1 実験設定

今回の実験では，提案手法の攻撃画像に対する防御可能性と，正規の入力データに対する分類精度の低下度合の検証を行なった．

まず，攻撃画像の準備では，手書き文字データセットの MMNIST を対象に，Cleverhans Library v3.1[17]で実装される FGSM 法と Szegedy 法を用いて，攻撃画像を生成した．あらかじめ分類器の学習に利用しなかったデータから生成された攻撃画像に対する防御可能性を確認するため，MNIST データセットのテストデータ 10,000 枚に対して，それぞれの敵対的サンプル生成手法を適用した．さらに，この中から攻撃により誤分類を起こす画像のみを利用するため，攻撃の元画像は正しく分類されるものの，攻撃画像は誤分類を起こすことを条件として抽出した．最終的にノイズ強度ε=0.10 を条件とした FGSM 法で生成した画像は 4,848 枚となった（図 2）．一方，誤分類させるクラスを指定する Szegedy 法では，誤分類先のクラスを正規のクラスから１シフトしたクラスとなることを条件に 8,148 枚を生成し，実験に利用した（図 3）．図 2 及び図 3 に生成された攻撃画像の一例を示しており，各画像のタイトル部分に，正規のクラスと誤分類されたクラスを記している．この２手法の間で生成された画像の枚数に差があるのは，FGSM 法で設定したノイズ強度が比較的小さくなるよう厳しめに設定したのに対して，Szegedy 法は最小化問題により攻撃画像を生成するものの，ノイズの強度に関する制約を与えていないことに起因する．また，正規の入力データに対する分類精度の低下度合の検証には，MNIST データセットのテストデータ 10,000 枚を利用した．

また，本実験を行うにあたり，利用する分類器には，３層の多層パーセプトロンを利用し，入力層は 784 ノードで，中間層はシグモイド関数を活性化関数とした 200 ノードの全結合層とし，最終層はソフトマックス関数を活性化関数とした 10 ノードの全結合層を利用した．学習には MNIST データセットの訓練データ 60,000 データを利用し，分類精度は 97.44%であった．

評価の指標には，攻撃画像に対する防御可能性としては，攻撃画像のうち，正規のクラスに分類できた画像の割合を利用する．また，正常な入力に対する精度の劣化は，正常画像全体をうち，正規のクラスに分類できた画像の割合を分類精度とし，こちらの指標は学習時の分類精度と同じものを用いた．いずれも 100%に近いほど，正しい分類ができていることを意味する．

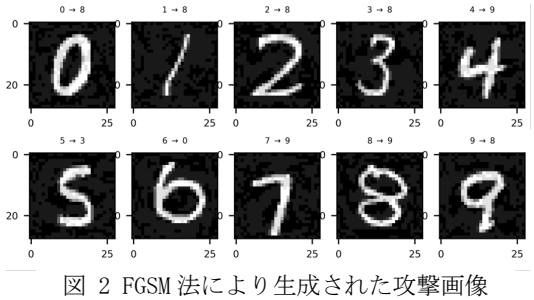

図 2 FGSM 法により生成された攻撃画像

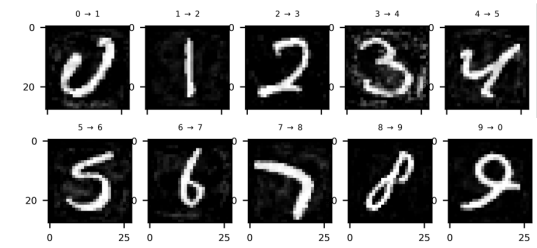

図 3 Szegedy 法により生成された攻撃画像

4.2 実験結果

実験結果を表 1 に示す．まず，攻撃画像に対する防御率では，FGSM 法では 83.46%，Szegedy 法では 95.08%が正規のクラスに分類される結果が得られた．どちらの攻撃手法に対しても，少なくとも 8 割以上の攻撃画像を正規のクラスへ分類できており，提案手法による防御の有効性が確認できた．また，正規の入力画像に対する分類精度の検証についても，NN の学習時は 97.44%であった分類精度に対して，提案手法による処理を追加であっても 96.34%を維持し，1.1%の低下にとどまった．この結果から，直交補空間方向への成分を除去することによる分類精度への影響は，十分に小さいことが確認できた．

表 1 提案手法の分類結果

	攻撃画像	正規の入力画像
	防御率	分類精度
FGSM 法	83.46%	96.34%
Szegedy 法	95.08%	

5. 結論

本稿では，学習データの多様体成分を抽出し，入力データから攻撃の要因となる多様体の直交成分を除去することにより，敵対的サンプルによる誤分類を防ぐ新たな防御手法を提案した．実験では，分類精度の劣化も十分に小さく，敵対的サンプルから防御できることを示した．発生メカニズムにおけるもうひとつの側面である重みの直交分解を組み合わせていくことにより，さらなる防御性能の向上も期待できることから，提案手法の発展版も評価を行い，今後報告していく予定である．なお，本稿における評価は，提案手法の有効性を示すための検証に留まるため，他手法との比較による本格的なベンチマークは今後の課題である．

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] 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).

[3] 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).

[4] 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”, https://www.digiarc.aist.go.jp/publication/aiqm/AIQM-Guideline-2.1.0.pdf, (最終アクセス日：2022/06/13)

[5] ENISA, “Cybersecurity Challenges in the Uptake of Artificial Intelligence in Autonomous Driving”, https://www.enisa.europa.eu/news/enisa-news/cybersecurity-challenges-in-the-uptake-of-artificial-intelligence-in-autonomous-driving, (最終アクセス日：2022/06/13)

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples”, arXiv preprint arXiv:1412.6572 (2014).

[7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards Deep Learning Models Resistant to Adversarial Attacks”, arXiv preprint arXiv: 1706.06083 (2017).

[8] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, “Distillation as a defense to adversarial perturbations against deep neural networks”, 2016 IEEE Symposium on Security and Privacy (SP), pp. 582-597 (2016)

[9] H. Tasaki, R. Lenz, and J. Chao, “Dimension estimation and topological manifold learning”, 2019 International Joint Conference on Neural Networks (IJCNN), pp.1–7 (2019).

[10] 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).

[11] H. Tasaki, Y. Kaneko, and J. Chao, “Curse of co-dimensionality: explaining adversarial examples by embedding geometry of data manifold”, International Conference on Pattern Recognition 2022 (発表予定)

[12] 足立 浩規, 平川 翼, 山下 隆義, 藤吉 弘亘, “[サーベイ論文] Adversarial Training”, 信学技報, vol. 121, no. 427, PRMU2021-73, pp. 78-90 (2022).

[13] S. Rifai, Y.N. Dauphin, P. Vincent, Y. Bengio, and X. Muller, “The manifold tangent classifier”, Advances in neural information processing systems, vol.24, pp.2294–2302 (2011).

[14] J. Wang, Geometric Structure of High-Dimensional Data and Dimensionality Reduction, Springer Berlin Heidelberg (2012).

[15] H.-M. Lu, Y. Fainman, and R. Hecht-Nielsen, “Image manifolds,” Applications of Artificial Neural Networks in Image Processing III, vol.3307, International Society for Optics and Photonics, pp.52–63 (1998).

[16] J. Lee, Introduction to Smooth Manifolds, Springer (2002).

[17] 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.