データ多様体の埋め込み幾何学に基づく新しい敵対攻撃法の提案

Novel Adversarial Attacks Based on Embedding Geometry of Data Manifolds

情報工学専攻　森田 匡博
Masahiro MORITA

概要

深層学習を利用した画像認識などでは，人間が知覚できないほど小さな摂動を加えて生成される敵対的サンプルによって，誤分類を引き起こすことが発見されている．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．本論文では，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しい敵対的サンプルの生成手法を提案し，これらの攻撃可能性についての評価を行う．

キーワード: 深層学習，敵対的サンプル，多様体仮説，データ多様体.

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，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摂動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].

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，確証ある理論は立てられていなかった．しかし最近，学習データが持つデータ多様体の埋め込み幾何学的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 [2].

本研究では，この発生メカニズムに基づいた新たな敵対的サンプルの生成手法に関して 3 つのアプローチを提案する．本手法は，データ変形への影響が少ない成分を抽出して摂動を生成するため，人に気づかれにくい摂動の生成が可能になる．本論文では，敵対的サンプルの生成手順を示した後，誤分類を起こす攻撃成功率の観点および生成される画像における摂動の可視性の観点から，提案手法の攻撃可能性を評価する．

2　敵対的サンプルに関する先行研究

2.1　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 [1]．学習は誤差を最小にすることを目指すが，FGSM では誤差を最大にして誤分類させることを目指している．敵対的サンプルは

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \epsilon \mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$$

で生成される．ここで，ϵ は摂動のサイズを制御するための定数，$\boldsymbol{\theta}$ はネットワークのパラメータ，$J(\boldsymbol{\theta}, \boldsymbol{x}, y)$ は誤差関数である．誤差関数の勾配を計算して，元のクラス y に対する誤差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，高速に生成することが可能となっている．また，2016 年には FGSM を標的型攻撃に拡張した手法が提案された [3]．敵対的サンプルは

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} - \epsilon \mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y_t))$$

で生成される．ターゲットクラス y_t の誤差関数を最小にする方向を摂動とし，敵対的サンプルを生成する．

2.2　データの多様体構造と埋め込み幾何学

ニューラルネットワークの分類対象となる画像データは，一般的に，画素数を次元とする高次元ベクトルで表現されるが，空間を充満するように存在するのではなく，限られた次元の部分多様体の上に存在する多様体仮説と呼ばれる性質をもつ．さらに，埋め込み幾何学とは，低次元部分多様体がどのように高次元空間に埋め込まれるかを記述する幾何学性質である．

2.3　敵対的サンプルの発生メカニズム

2022 年に田崎らはデータ多様体の埋め込み構造に基づく発生メカニズムを提案した [2]．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部分多様体構造を持つ．ここで 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 $n+1$ 次元の射影空間 $\mathcal{S}$ を考えると，データ多様体 M は $\mathcal{M} = \{\mathbf{x} = (\boldsymbol{x}^\top, 1)^\top | \ \boldsymbol{x} \in M\}$，重みは $\mathbf{w} = (\boldsymbol{w}^\top, \theta)^\top$ となる．ここで，$\mathcal{M}$ 上のある点 $\mathbf{x}$ における接空間 $T_{\mathbf{x}}\mathcal{M}$ とそれに直交する空間である直交補空間 $T_{\mathbf{x}}^{\perp}\mathcal{M}$ への直交分解

$$T_{\mathbf{x}}\mathcal{S} = T_{\mathbf{x}}\mathcal{M} \oplus T_{\mathbf{x}}^{\perp}\mathcal{M}$$

が成立する．これにより，点 $\mathbf{x}$, 重み $\mathbf{w}$, 内積 $\mathbf{w}^\top\mathbf{x}$ はそれ

ぞれ $\mathbf{x} = \mathbf{x}_{\mathcal{M}} + \mathbf{x}_{\mathcal{M}}^{\perp}, \mathbf{w} = \mathbf{w}_{\mathcal{M}} + \mathbf{w}_{\mathcal{M}}^{\perp}, \mathbf{w}^{\top}\mathbf{x} = \mathbf{w}_{\mathcal{M}}^{\top}\mathbf{x}_{\mathcal{M}} + (\mathbf{w}_{\mathcal{M}}^{\perp})^{\top}\mathbf{x}_{\mathcal{M}}^{\perp}$ と表せ，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．ここで，摂動 $\mathbf{r}$ が加わった敵対的サンプル $\tilde{\mathbf{x}}$ がニューラルネットワークに入力された際の重みとの内積は

$$\mathbf{w}^{\top}\tilde{\mathbf{x}} = \mathbf{w}_{\mathcal{M}}^{\top}\mathbf{x}_{\mathcal{M}} + \mathbf{w}_{\mathcal{M}}^{\top}\mathbf{r}_{\mathcal{M}} + (\mathbf{w}_{\mathcal{M}}^{\perp})^{\top}\mathbf{r}_{\mathcal{M}}^{\perp}$$

と表せる．多様体方向の成分 $\mathbf{r}_{\mathcal{M}}$ 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，誤分類は直交補空間方向の摂動 $\mathbf{r}_{\mathcal{M}}^{\perp}$ によって引き起こされる．

3 提案手法

本論文では，学習データが持つデータ多様体の埋め込み構造に基づく，データ多様体の直交補空間方向に対応したニューラルネットワークの重みを活用した敵対的サンプルの生成手法を提案する．敵対的サンプルは以下の手順に従って生成されるものとする．

Step 1. **射影空間への埋め込み**

学習データ $\boldsymbol{x}_i$ に対するデータ行列を $\boldsymbol{X} = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_N]$ とし，1 層目の中間層における各ニューロンの重み $\boldsymbol{w}_i$ に対する重み行列を $\boldsymbol{W} = [\boldsymbol{w}_1 \cdots \boldsymbol{w}_h]$ とする．これらを射影空間に埋め込み，データ点を $\mathbf{x}_i = (\boldsymbol{x}_i^{\top}, 1)^{\top}$, 重みを $\mathbf{w}_i = (\boldsymbol{w}_i^{\top}, \theta_i)^{\top}$ とするとき，$\mathbf{W} = [\mathbf{w}_1 \cdots \mathbf{w}_h]$ となる．N は学習に用いるデータ数，h は 1 層目の中間層のニューロン数である．

Step 2. **近傍分割**

入力データ $\mathbf{x}$ からの距離が最短となる点を探索し，最短距離の点から距離が近い順に k 点を取得し，$\mathbf{x}$ の近傍とする．

Step 3. **接空間の基底の算出**

得られた近傍に対して，局所的に主成分分析（PCA）を適用する手法である LPCA を利用し，累積寄与率が定めた閾値 φ 以上になる次元 d を接空間の次元として推定する．さらに，主成分分析により求められる主成分または固有ベクトルを $\boldsymbol{u}_i$ とするとき，固有値が大きい順に対応する固有ベクトル $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d$ による行列は $\boldsymbol{U}_{\mathcal{M}} = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_d]$ であって，これは接空間 $T_{\mathbf{x}}\mathcal{M}$ の直交基底となる．

Step 4. **摂動の計算**

算出された接空間成分を元に，摂動 $\mathbf{r}$ を求める．詳細は 3.1 節から 3.3 節で述べる．

Step 5. **敵対的サンプルの生成**

敵対的サンプル $\tilde{\mathbf{x}}$ を $\tilde{\mathbf{x}} = \mathbf{x} + \epsilon\frac{\mathbf{r}}{\|\mathbf{r}\|}$ で求め，アフィン空間上に戻し敵対的サンプル $\tilde{\boldsymbol{x}}$ を得る．

Step 6. **クリップ処理**

クリップ処理により，敵対的サンプルの画素値を $0.0 \sim 1.0$ の範囲に抑える．

3.1 重みの直交補空間方向をすべて利用した攻撃

ニューラルネットワーク内の内積計算式における第 3 項 $(\mathbf{w}_{\mathcal{M}}^{\perp})^{\top}\mathbf{r}_{\mathcal{M}}^{\perp}$ を最大化させるために，摂動を重みの直交補空間方向とする手法が考えられる．摂動 $\mathbf{r}$ は

$$\mathbf{r} = \boldsymbol{U}_{\mathcal{M}}^{\perp}(\boldsymbol{U}_{\mathcal{M}}^{\perp})^{\top}\mathbf{W}\,\mathbf{1}_h = \mathbf{W}_{\mathcal{M}}^{\perp}\,\mathbf{1}_h$$

で求められる．$\mathbf{W}_{\mathcal{M}}^{\perp}$ 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，$\mathbf{W}_{\mathcal{M}}^{\perp} = \mathbf{W} - \mathbf{W}_{\mathcal{M}} = \mathbf{W} - \boldsymbol{U}_{\mathcal{M}}(\boldsymbol{U}_{\mathcal{M}})^{\top}\mathbf{W}$ と変形できる．$\mathbf{1}_h$ は全要素が 1 の h 次元列ベクトルである．以下，本攻撃手法を攻撃法 I とする．

3.2 中間層の出力を利用した攻撃

敵対的サンプルによる誤分類は，ニューラルネットワークの中間層における活性化関数の出力が変化することによって生じる．事前の解析にて，重みと正常画像の内積をシグモイド関数に与えた結果，0.0 あるいは 1.0 付近に集中していたため，ここでは，中間層の出力が反転するような重みの直交補空間方向を用いて摂動を生成する手法について述べる．

まず, ある中間層のニューロンの重み $\mathbf{w}_i$ とそのニューロンの入力 $\mathbf{x}$ との内積値を求め，シグモイド関数 f に入力として与え，出力 $y_i = f((\mathbf{w}_i)^{\top}\mathbf{x})$ を得る．そして，出力を成分にもつベクトルが $\boldsymbol{y} = [y_1 \cdots y_h]^{\top}$ で表せる．次に，$\boldsymbol{y}$ を 3 つの範囲に分割し，それぞれの範囲に当てはまる y_i を反転させるような重みの直交補空間ベクトルを利用する．本手法では，0.2 以下の範囲を取る重みの直交補空間ベクトルの部分和ベクトルを $\mathbf{r}_{+}$，0.8 以上の範囲を取る重みの直交補空間成分の部分和ベクトルを $\mathbf{r}_{\text{-}}$ とする．この 2 つのベクトルを用いて摂動 $\mathbf{r}$ を求めるが，ベクトルという都合上互いに打ち消しあってしまう可能性があるため，以下の 3 パターンを考慮する必要がある．

$$\mathbf{r} = \begin{cases} \mathbf{r}_{+} & (1) \\ \mathbf{r}_{+} - \mathbf{r}_{\text{-}} & (2) \\ -\mathbf{r}_{\text{-}} & (3) \end{cases}$$

以下，(1), (2), (3) の各摂動によって敵対的サンプルを生成する手法をそれぞれ攻撃法 II-(1), 攻撃法 II-(2), 攻撃法 II-(3) とする．

3.3 ターゲットクラス画像の中間層の出力を利用した攻撃

最後に，ターゲットクラス画像の中間層の出力を利用した攻撃について説明する．

まず，重み $\mathbf{w}_i$ と，訓練データ内においてニューラル

ネットワークのターゲットクラス t への信頼度が一番高い画像 $\mathbf{x}_t$ の内積を計算し，シグモイド関数 f に入力として与え，出力 $y_{t_i} = f((\mathbf{w}_i)^\top \mathbf{x}_t)$ を得る．そして，出力を成分にもつベクトルが $\boldsymbol{y}_t = [y_{t_1} \ \cdots \ y_{t_h}]^\top$ で表せる．同様に，クラス $o(\neq t)$ の元画像 $\mathbf{x}_o$ を用いて $\boldsymbol{y}_o = [y_{o_1} \ \cdots \ y_{o_h}]^\top$ を求める．次に，中間層の出力ベクトルの差 $\boldsymbol{d}(t,o)$ を

$$\boldsymbol{d}(t,o) = \boldsymbol{y}_t - \boldsymbol{y}_o$$

で求める．ここで，$\boldsymbol{d}(t,o)$ の i 番目の要素が 0.8 以上なら i 番目のニューロンを増加させる重みの直交補空間成分の部分和ベクトルを $\mathbf{r}_+$，-0.8 以下なら減少させる重みの直交補空間成分の部分和ベクトルを $\mathbf{r}_-$ とする．最後に，それらを合わせたベクトルを摂動 $\mathbf{r} = \mathbf{r}_+ - \mathbf{r}_-$ とする．以下，本攻撃手法を攻撃法 Ⅲ とする．

4 実験

敵対的サンプルの発生メカニズムに基づく，データ多様体 M の直交補空間 $T_{\mathbf{x}}^{\perp}\mathcal{M}$ を利用した提案手法の有効性を検証する．

4.1 実験設定

本実験では，MNIST 内の訓練データ中の 1 万枚と，Fashion-MNIST 内の訓練データ中の 1 万枚の 2 種類のデータセットを対象に攻撃画像を生成する．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採用し，入力層は 784 ニューロン，中間層はシグモイド関数を活性化関数とする 200 ニューロンの全結合層を 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 ニューロンの全結合層で構成した．なお，分類精度は MNIST で 97.92%，Fashion-MNIST で 88.24% である．攻撃の生成においては，1 つの近傍に属する点の個数を 150 点，接空間の推定における PCA の累積寄与率に対する閾値 φ を $\varphi = 0.99$ とする．

4.2 実験結果

4.2.1 攻撃成功率

FGSM と提案手法によって生成された敵対的サンプルの誤分類率（攻撃成功率）を表 1 および表 2 に示す．攻撃法 I による攻撃は，あまり攻撃成功率が伸びず FGSM より一回り低い攻撃成功率となったが，攻撃法 Ⅱ による攻撃は，摂動サイズ ϵ を大きくした際に FGSM と同程度の攻撃成功率であった．

4.2.2 各空間における摂動の長さの比較

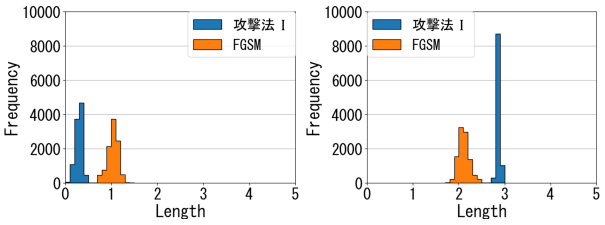
$\epsilon = 0.1$ での FGSM と攻撃法 I における，摂動の接空間成分の長さ $\|\mathbf{r}_{\mathcal{M}}\|$ と直交補空間成分の長さ $\|\mathbf{r}_{\mathcal{M}}^{\perp}\|$ のヒストグラムを図 1 および図 2 に示す．これらより，(a) のヒストグラムでは，FGSM には接空間方向の成分が多く含まれていることがわかった．一方で，(b) のヒストグラムでは，提案手法には直交補空間方向の成分が多く含まれることがわかった．これらより，FGSM が生成する摂動には，誤分類を起こすために多様体成分が多く含まれており，人間に気づきにくいと定義される敵対的サンプルとは異なる攻撃画像が含まれていると考えられる．

表 1 各攻撃手法による攻撃成功率 (MNIST)

攻撃手法	攻撃成功率 (%)			
$\epsilon =$	0.05	0.10	0.15	0.20
FGSM	53.82	89.81	97.83	99.42
攻撃法 I	19.05	47.75	63.33	73.66
攻撃法 Ⅱ-(1)	33.31	83.04	96.90	99.68
攻撃法 Ⅱ-(2)	20.03	68.40	88.47	95.78
攻撃法 Ⅱ-(3)	5.07	34.23	60.90	77.41

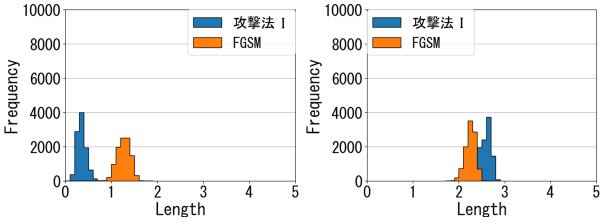
表 2 各攻撃手法による攻撃成功率 (Fashion-MNIST)

攻撃手法	攻撃成功率 (%)			
$\epsilon =$	0.05	0.10	0.15	0.20
FGSM	68.19	90.76	93.77	94.37
攻撃法 I	17.28	35.30	50.92	60.49
攻撃法 Ⅱ-(1)	17.89	54.92	82.65	92.71
攻撃法 Ⅱ-(2)	16.40	52.72	82.69	93.41
攻撃法 Ⅱ-(3)	14.84	47.37	73.48	89.35



(a) 接空間成分の長さ $\|\mathbf{r}_{\mathcal{M}}\|$ **(b)** 直交補空間成分の長さ $\|\mathbf{r}_{\mathcal{M}}^{\perp}\|$

図 1 各空間における摂動の長さ (MNIST)



(a) 接空間成分の長さ $\|\mathbf{r}_{\mathcal{M}}\|$ **(b)** 直交補空間成分の長さ $\|\mathbf{r}_{\mathcal{M}}^{\perp}\|$

図 2 各空間における摂動の長さ (Fashion-MNIST)

4.2.3 生成される攻撃画像の比較

図 3 および図 4 に示すように，$\epsilon = 0.2$ における FGSM と各攻撃法によって生成された敵対的サンプルの描画を行った．左列から元画像，(a) FGSM，(b) 攻撃法 I，(c) 攻撃法 Ⅱ − (2) である．これらより，(a) の FGSM が生成する画像では，被写体そのものを変化させて誤分類を起こすような摂動を生成していることがわかる．一方で，(b) および (c) の提案手法で生成された攻撃画像には，字体の変形ではなく背景雑音が含まれる傾向にあった．

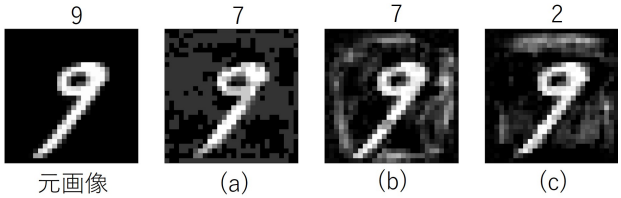


図 3 MNIST に対する敵対的サンプル (左列から元画像, (a) FGSM, (b) 攻撃法 I, (c) 攻撃法 II-(2))

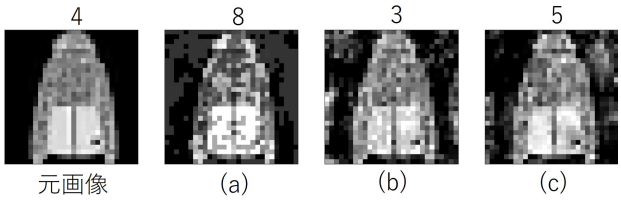


図 4 Fashion-MNIST に対する敵対的サンプル (左列から元画像, (a) FGSM, (b) 攻撃法 I, (c) 攻撃法 II-(2))

表 3 Targeted FGSM の各ターゲットクラスへの誤分類率 (MNIST)

o \ t	0	1	2	3	4	5	6	7	8	9
0	-	0.20	91.61	75.62	27.57	73.13	96.8	93.01	51.55	81.42
1	19.79	-	100.00	99.82	7.45	43.66	99.82	99.91	99.82	86.34
2	31.48	19.37	-	89.81	16.75	71.75	94.35	91.02	83.85	58.02
3	51.07	3.88	96.22	-	9.30	83.91	84.30	78.59	79.26	86.24
4	8.98	1.02	80.71	88.78	-	73.57	99.80	99.29	65.82	97.24
5	23.17	1.74	97.91	93.51	11.12	-	94.67	93.97	45.65	79.72
6	55.72	0.00	98.42	83.63	32.45	67.85	-	91.42	69.43	66.77
7	17.94	2.71	96.82	89.81	20.00	73.64	82.71	-	83.18	83.18
8	22.25	2.65	98.52	86.23	13.56	64.09	87.29	94.28	-	79.98
9	4.91	0.61	79.14	78.94	48.47	65.44	99.39	95.30	20.04	-

表 4 攻撃法 III の各ターゲットクラスへの誤分類率 (MNIST)

o \ t	0	1	2	3	4	5	6	7	8	9
0	-	0.00	100.00	92.41	1.30	98.20	100.00	100.00	93.11	73.83
1	38.78	-	100.00	92.64	1.15	89.26	99.91	100.00	88.55	90.24
2	66.20	0.00	-	99.80	0.81	76.79	100.00	100.00	90.11	81.13
3	49.71	0.00	100.00	-	0.48	92.64	100.00	100.00	83.43	80.14
4	10.10	0.00	99.80	95.92	-	93.37	100.00	100.00	78.67	60.51
5	24.68	0.00	100.00	98.96	1.39	-	100.00	100.00	69.52	21.55
6	47.04	0.00	100.00	97.44	2.47	78.21	-	100.00	85.90	63.61
7	3.18	0.00	100.00	96.54	0.09	72.24	100.00	-	72.90	44.11
8	60.17	0.21	100.00	99.36	0.42	68.86	99.89	100.00	-	32.94
9	1.43	0.00	99.80	92.64	1.53	83.64	100.00	100.00	60.53	-

4.2.4 ターゲットクラスへの誤分類率の比較

$\epsilon = 0.2$ の状況下で，Targeted FGSM と攻撃法 III での，各元クラス o とターゲットクラス t における誤分類率を表 3 から表 6 に示す．これらより，MNIST および Fashion-MNIST では既存手法と比較して，誤分類率は同等以上の結果となった．

5 結論と今後の課題

先行研究で提案された，データ多様体の直交補空間方向の摂動により誤分類が引き起こされるという発生メカニズムに基づき，既存手法とは全く異なる新しい敵対的サンプルの生成手法を提案した．提案手法は，非標的型攻撃と，指定した任意のクラスへ誤分類させる標的型攻撃を実現し，その有効性を確認した．また，データの変形に影響が少ない直交補空間方向のみを摂動とすることにより，元画像の見た目への影響を抑えた敵対的サンプルが生成されることを確認した．今後の課題として，生成される摂動の可視性の評価方法の確立やその基準に従った攻撃方法の検討が挙げられる．

表 5 Targeted FGSM の各ターゲットクラスへの誤分類率 (Fashion-MNIST)

o \ t	0	1	2	3	4	5	6	7	8	9
0	-	48.83	71.23	74.52	53.29	24.10	99.79	12.10	100.00	38.00
1	99.51	-	79.07	83.74	67.67	24.73	63.00	57.45	99.81	82.96
2	44.59	26.97	-	40.35	92.03	41.04	99.90	6.50	100.00	25.20
3	100.00	97.55	91.07	-	93.23	45.04	95.39	30.03	100.00	84.30
4	40.66	36.34	96.30	27.31	-	29.67	100.00	5.13	100.00	15.61
5	39.43	18.30	71.59	36.91	26.59	-	80.59	70.37	99.90	96.76
6	65.82	49.85	93.44	56.02	76.69	33.20	-	9.30	100.00	37.22
7	14.09	4.40	58.41	3.23	5.68	100.00	85.13	-	100.00	99.61
8	67.17	30.10	94.85	56.16	54.14	74.75	99.09	45.86	-	73.94
9	6.70	2.80	2.30	2.10	1.10	96.80	15.90	15.30	95.60	-

表 6 攻撃法 III の各ターゲットクラスへの誤分類率 (Fashion-MNIST)

o \ t	0	1	2	3	4	5	6	7	8	9
0	-	97.98	98.09	97.13	44.69	6.05	92.57	15.50	100.00	57.86
1	99.81	-	99.32	79.84	32.91	34.57	56.18	60.27	100.00	91.24
2	88.68	96.16	-	84.45	52.36	3.44	89.57	10.04	100.00	42.42
3	99.71	86.36	99.21	-	57.31	24.44	86.85	47.60	100.00	95.00
4	77.21	94.15	98.67	60.78	-	2.26	91.07	10.16	100.00	36.96
5	99.70	91.81	90.29	83.62	10.01	-	99.49	31.85	99.39	86.15
6	87.27	96.28	93.93	90.60	45.84	5.09	-	17.34	100.00	59.75
7	99.80	91.39	99.71	94.42	30.92	76.32	100.00	-	100.00	96.28
8	76.97	89.29	98.28	84.44	31.82	43.94	99.09	26.16	-	76.26
9	98.70	68.30	94.20	46.90	33.00	97.20	98.90	21.30	100.00	-

謝辞

本研究を進めるにあたり，適切な御指導を頂いた趙 晋輝 教授と，情報通信工学研究室 博士後期課程の田崎 元さんに，感謝の意を表します．

関連発表

1. 森田 匡博, 田崎 元, 趙 晋輝, “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,” 第 21 回情報科学技術フォーラム, 2022 年 9 月 14 日.(FIT 奨励賞受賞)
2. 森田 匡博, 田崎 元, 趙 晋輝, “データ多様体の埋め込み幾何学に基づく新しい敵対攻撃法の提案,” 第 49 回 IBISML 研究会, 2023 年 3 月 2 日.(発表予定)

参考文献

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

[2] H. Tasaki, Y. Kaneko, and J. Chao, “Curse of co-dimensionality: Explaining adversarial examples by embedding geometry of data manifold,” 2022 26th International Conference on Pattern Recognition(ICPR)IEEE, pp.2364–2370, 2022.

[3] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial machine learning at scale,” arXiv preprint arXiv:1611.01236, 2016.