Vol.2020-SE-205 No.1
2020/7/2

敵対的なセマンティック・ノイズの実行時検知

中島 震[1,a)]

概要：機械学習ソフトウェアは敵対データが入力されると素朴な期待に反する振舞いを示す．不具合の一種と考えられ，防御や検知法の研究が精力的に進められているが，一般的な解決策は未だ見出されていない．本稿は，選択的に敵対的なセマンティック・ノイズに対して汎化性能が劣るように訓練した学習モデルを用いて，予測齟齬に基づく敵対データ検知の方法を提案する．劣った汎化性能に注目することの基本的な見方を整理し実験の結果を報告する．

キーワード：深層ニューラル・ネットワーク，ロバスト性，予測の齟齬，ゴースト成分

Detecting Adversarial Semantic Noises at Runtime

Shin Nakajima[1,a)]

Abstract: Machine learning software exhibits undesirable behavior when input data are adversarial examples. Although lots of methods have been proposed and studies, no general solution for adversarial defenses or detections is discovered. This paper proposes a new adversarial detection method, with the prediction inconsistency approach, to employ machine learning models so trained as to show poor generalization performance selectively for adversarial semantic noises. The discussion includes the rationale of the approach and some results of the experiments.

Keywords: Deep Neural Networks, Robustness, Prediction Inconsistency, Ghost Features

1. はじめに

敵対データ（Adversarial Examples）は，目視による直感的な結果を仕様とする時，これに反する予測を導く [17][19]．敵対データが深層ニューラル・ネットワーク（DNN）の不適切な振舞いを導く理由として，その学習方式に本質的な原因があると考えられた [5]．その後，画像に追加した微弱なノイズ（敵対擾乱）だけでなく，画像化される前の実世界の物体への加工によって敵対データが作れる [3] など，さまざまな敵対データ作成方法が考案された [20]．

敵対データがさまざまな興味深い性質を持つことがわかっている [7]．訓練済み学習モデルの情報を活用して敵対データを生成する方法をホワイトボックス攻撃と呼ぶ．このような情報を利用できない状況で敵対データを生成するブラックボックス攻撃も可能なことが実験で示された．また，訓練データセット間，学習モデル間，学習方式間，で敵対データが転用可能である．

敵対データは一種の不具合を生じることから，その防御あるいは検知法への関心が高い [7][15]．防御の方法は，入力データが持つ敵対性を無効化するように，誤予測を生じないように DNN 学習モデルを訓練することに関心がある．しかし，多様な敵対データを防御できる一般性の高い解決法は見つかっていない．そこで，最近，入力データが敵対擾乱を含むかを調べる検知の方法に関心が移ってきた [2]．検知機構を通常の予測推論プログラムと同時に作動させるアーキテクチャを採用するものである．

本稿は，入力データが敵対性を持つか否かを予測推論結果から推定する方法を提案し，簡単な実験結果を報告する．ここで用いた予測齟齬（Prediction Inconsistency）の方法

1 国立情報学研究所
NII, Chiyoda-Ku, Tokyo 101—8430, Japan
a) nkjm@nii.ac.jp

ⓒ 2020 Information Processing Society of Japan

は，異なる汎化性能を持つ複数の DNN モデル[*1]を同時に作動させ，予測推論の結果が合致しないことを利用する．訓練データが性質の異なる成分を持つという仮説のもと，成分ごとに異なる汎化特性を示す DNN モデルを構築する，という方法である．

以下，第 2 節は敵対データに関連する技術背景を紹介する．第 3 節は訓練データが性質の異なる成分を持つという仮説を示し，異なる成分が汎化に影響するかを確認する実験を行う．第 4 節は第 3 節の結果を踏まえて，敵対データを実行時分離する評価実験を行う．第 5 節で実験方法ならびに実験結果を考察し，第 6 節で本稿を終える．

2. 敵対データ

2.1 予測推論の性能

一般に，機械学習に期待される「気が利く」という特徴は，訓練学習時のデータセット XS が含まない入力データに対して妥当な予測結果を返すことにある．分類学習タスクを考えて，正確性とロバスト性の 2 つを整理する．

今，$Prob(M, \vec{a}, c)$ を DNN モデル M が入力データ $\vec{a}$ をカテゴリ c に分類する確率（予測の確からしさ）とする．素朴には，最大確率となる分類カテゴリ c^* が良い指標になる．正解タグを既知とする評価用のデータセット $ES = \{\langle \vec{x}^m,\ t^m \rangle\}$ を与え，$\vec{x}^m$ に対する c^* が正解タグ t^m に一致する頻度（正解率）を調べれば良い．この正解率に着目した性能を正確性と呼ぶ．訓練学習時の XS に対して良い正確性を示す一方で ES の正確性が悪いと，汎化誤差が大きいので汎化性能が劣る．

予測結果のロバスト性は，$\|\cdot\|$ を適切なメトリックスとする時，既知のデータ $\vec{x}^n$ から少しズレた未知のデータ ($\|\vec{x} - \vec{x}^n\| \leq \delta$) に対して，その予測結果 $\mathcal{I}_f(\vec{x})$ が大きく変わらないことをいう．ϵ を微小値として，$\|\mathcal{I}_f(\vec{x}) - \mathcal{I}_f(\vec{x}^n)\| \leq \epsilon$ である．大きな δ に対して，ϵ が適切な小さい値に収まる時，ロバスト性が良く，未知データからなる ES に対して，期待を外さない（驚きのない）予測結果を示す．これは素朴な意味での汎化性能が良いことである．分類問題の場合，簡単には，確率最大の分類カテゴリ c^* が同一になるような最大の δ_c をロバスト半径と考えれば良い．つまり，δ_c を超えると別カテゴリに分類される．

本稿で考える敵対データ [17] は，目視上のイメージ空間の距離が小さいにも関わらず，素朴な期待に合わない予測結果（分類結果）を返すことから，ロバスト性の考え方に大きな影響を与えた．入力データが敵対性を持ち誤予測を生じると見做す．敵対性の可能性を考慮しても，素朴に期待通りの予測結果になること，つまり，誤予測が生じないことを敵対ロバスト性（Adversarial Robustness）と呼ぶ．基準データから擾乱によって少しズレた入力データに対しても分類結果が一致するような最大の δ_A を敵対ロバスト半径とする．敵対的な結果を導く敵対擾乱が微小であれば，$\delta_A < \delta_c$ となる可能性が大きく，実際，実験で確認されている．つまり，ロバスト半径内に敵対データが存在することになり，ロバスト性の定義と整合しない．

[*1] 本稿では訓練済み DNN 学習モデルを単に DNN モデルと呼ぶ．

(a) 元画像　(b) 敵対擾乱　(c) 歪みとランダムノイズ

図 1 セマンティック・ノイズ

さて，敵対データが誤予測するとしても，その予測結果は確からしさを伴う確率として表される．この確率が大きいとしても，それは，素朴な期待に反する分類結果が「確からしい」ことを示すものであって，通常のデータに対する「期待通り」の分類結果の確からしさとはニュアンスが異なる．そこで，このような敵対的な予測確からしさを，「敵対的な正確性」と呼ぶことにする．

2.2 セマンティック・ノイズ

2.2.1 いくつかの例

敵対データは多くの研究者の関心を集め，さまざまな敵対データ生成法が提案されてきた [20]．以下，敵対データ生成の簡単な方法をもとに基本的な考え方を説明する．

まず，いくつかの目視イメージを図 1 に示した．基準となる元データ $\vec{x}^m$（図 1(a)），$\vec{x}^m$ にランダム・ノイズ $\vec{\Delta}^m$ を付加したデータ（図 1(c) の最右上下 2 枚），さらに，以下に示す自動生成法でセマンティック・ノイズを追加したデータ（図 1(b) および図 1(c) の左上下 2 枚）である．本稿では，敵対データのように予測推論結果に影響を与えることから，ランダム・ノイズに比してセマンティック・ノイズと呼ぶ．

2.2.2 自動生成法

セマンティック・ノイズを伴うデータは，L-BFGS 法 [17] を利用して自動生成することができる．元データ $\vec{x}_S$ をラベル t_T に誤分類させる敵対データは制約条件付きの最適化問題の解 $\vec{x}^*$ として得られる．W^* を，学習モデル $\vec{y}$ に関する訓練済み学習パラメータの集まりとして，

$$A_\lambda(W^*; \vec{x}_S, t_T, \vec{x}) \ = \ \ell(\vec{y}(W^*;\ \vec{x}),\ t_T) + \lambda \ell'(\vec{x},\ \vec{x}_S)$$

に対して，

$$\vec{x}^* = \underset{\vec{x}}{argmin}\ A_\lambda(W^*; \vec{x}_S, t_T, \vec{x})$$

$$\textbf{s.t.}\ \ 0 \leq x_j \leq 1 \ \wedge\ \Psi(\vec{x}, \vec{x}_S)$$

によって求める．関数 A_λ の定義式中，第 1 項は $\vec{x}$ に対する予測結果が与えられた教師タグ t_T を再現すること，第 2 項は $\vec{x}$ が指定のデータ $\vec{x}_S$ に「近い」ことを表す．この

時，上式で，うまくハイパーパラメータ λ の値を選ぶと，元データ $\vec{x}_S$ からの差が小さくなるような $\vec{x}^*$ を得ることができる．その結果，$\vec{x}^*$ は，人間の眼には $\vec{x}_S$ に見える一方で，t_T に分類させるようなセマンティック・ノイズを含み敵対データである．

文献 [17] の方法は，$\Psi(\vec{x}, \vec{x}_S)$ を論理真とした場合に相当する．上記では，式 $\Psi(\vec{x}, \vec{x}_S)$ によって，生成するデータ $\vec{x}^*$ に補助的な制約条件を付加できるようにした．また，ℓ' として L_2 ノルムに基づく損失関数を用いた．図 1(b) の左上下の 2 枚は補助的な制約条件のない場合であるが，図 1(b) の右上は額縁にみえるセマンティック・ノイズ．右下は文献 [3] が示した黒テープ形状の敵対データをエミュレートしたセマンティック・ノイズである．

2.2.3 歪みデータ

次に，最適化問題に与えるパラメータを変更して，試験データセット TS の要素 $\langle \vec{x}^m, t^m \rangle$ に対して，セマンティック・ノイズを挿入したデータ $\vec{x}^{(m)*}$ を得るようにする．

$$\vec{x}^{(m)*} = \underset{\vec{x}}{argmin}\ A_\lambda(W^*; \vec{x}^m, t^m, \vec{x})$$

$\vec{x}^{(m)*}$ は元の正解タグ t^m を再現するが，$\vec{x}^m$ から相対的な歪みがあると考えて良い．λ が微小値の場合，相対的な歪みの度合いが大きい．λ が適切な大きさの時，$\vec{x}^{(m)*}$ はセマンティック・ノイズの影響が小さく，$\vec{x}^m$ とほぼ一致する．

2.2.4 敵対データとノルム

図 1(c) の左上下の 2 枚は補助的な制約条件を付加しての歪みデータ生成例である．目視上，図 1(b) の右上下 2 枚と区別がつかないが，セマンティック・ノイズの現れが敵対擾乱であるか，歪みのノイズであるか，大きく異なる．

ここで多次元ベクター・データ間距離について整理しておく．$1 \leq p < \infty$ に対して p ノルム L_p は $\| \vec{x} \|_p \equiv (\sum |x_j|^p)^{1/p}$ であり，$p = 2$ の時，ユークリッド空間の距離に一致する．また，無限ノルム L_∞ は $\| \vec{x} \|_\infty \equiv max(|x_j|)$ である．ゼロノルム L_0 は non-zero 成分の個数を表す．

ℓ' として用いるノルムによって，イメージ空間での元データからの距離が異なり．その結果，異なる性質を持つ敵対データを生成することができる．既存の攻撃法は特定のノルムに基づく．たとえば，FGSM [5] は L_∞，BIM は L_∞，DeepFool [11] は L_2，JSMA は L_0，Carlini&Wagner [1] は L_2 と L_∞ と L_0 である．

2.3 敵対データの防御と検知

汎用の防御法，多様な攻撃を無効化する方法は，技術的に困難である．敵対的な攻撃をセキュリティ問題（例えば [15]）と考える場合，攻撃者は防御あるいは検知の方法を知ることで新たな攻撃法を考案できる [2]．

機械学習の自然な発想としては，敵対データに対する耐性（敵対ロバスト性）が改善するような DNN モデルを得れば良い．既知の敵対データを訓練時に用いるデータ補完の

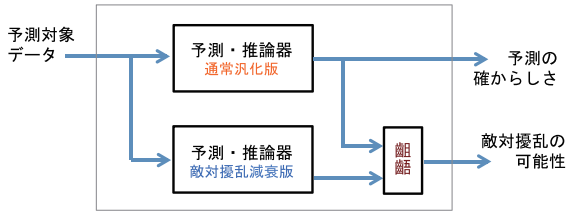

図 2 アーキテクチャ

アプローチがある．補完した敵対データ生成法の特徴を考慮して訓練学習で用いる誤差関数（最適化の目的関数）を工夫する勾配マスキング（Gradient Masking）[5] などの方法を敵対訓練（Adversarial Training）と総称する．また，これらの方法を整理した敵対性学習（Adversarial Learning）[10] が，L_∞ ノルムの場合に検討されている．これらの防御方法を実現したとしても，想定範囲外の攻撃に対して脆弱さが残る．

特定の攻撃法を想定しない汎用な方法として，イメージ空間のノイズ除去（De-noising），ノイズによる微小変化を平準化するように入力データを前処理する入力データ変換（Input Transformation）がある．一般に，入力データの情報が減ることから，クリーン・データに対する予測性能に悪い影響を及ぼすことが多い．一方，出力の予測結果を調べる予測齟齬（Prediction Inconsistency）や，DNN モデルの中間層から得られる情報を統計的な方法で活用するアプローチがある．これらの方法では，クリーン・データを敵対データと判定する誤検出（Spurious Alarms）ならびに敵対データがすり抜ける見落とし（Overlook）が生じる確率が問題となる．誤検出ならびに見落としの両方を小さくしたいが一般には両立しない．

3. 選択的な汎化

3.1 モデル・キャパシティ

敵対データの原因を DNN モデルの予測推論機構に内在すると考えるのではなく，訓練データセットの要素である多次元データに着目する研究がある．

通常のデータ $\vec{x}^n$ は，対応する正解タグ t^n を真に再現するのに不可欠な成分（純化成分と呼ぶ）に加えて，存在しなくても正解タグを再現し得る付加的な成分（ゴースト成分と呼ぶ）を暗に含むと考える．そして，DNN モデルは純化成分ならびにゴースト成分すべてを「適切に」汎化するように機能する．ゴースト成分から得られる情報は元データ $\vec{x}^n$ と異なる入力データに対しても，「適切に」元データに対応する正解タグ t^n を再現する．すなわち，敵対性が生まれる原因は，DNN の学習方式に問題があるわけではない．原因は訓練データにあり，DNN モデルの汎化機構が適切に作用することで入力データが含むセマンティック・ノイズを敵対的な結果に導いた，と論じる [9]．

ここで，文献 [5] による「パンダ」と「手長ザル」の例に

ついて，敵対データ生成とゴースト成分の関係を想像してみる．元データの「パンダ」の確からしさは 57.7%で，敵対データの「手長ザル」の確からしさは 99.3%である．元データの確からしさが小さい「パンダ」データは純化成分の割合が低いと考えて良いだろう．そこで，元データのゴースト成分と敵対擾乱とが重なり合い確からしさの高い「手長ザル」敵対データを得たと論じることができる．実際，元データの正確性が低いほど，敵対擾乱を加えたデータの敵対的な正確性が高くなることが経験的にわかっている．

一方，敵対データの影響を被る DNN モデルは許容度（モデル・キャパシティ）が小さいことに原因があるとする見方がある [14]．ヒトは目視によって敵対データに欺かれることはない．十分なモデル・キャパシティ[*2]があれば敵対データに対する誤予測が生じないと論じる．つまり，通常のデータ $\vec{x}^n$ が含む全ての成分を活用して「適切に」汎化するので，直感通りの予測結果が得られる，と考える．

上記の 2 つの議論は，モデル・キャパシティの違いによって，汎化の結果が異なることを示唆する．文献 [9] は DNN モデルのモデル・キャパシティは十分でなく敵対データの影響を被る程度の許容度であることを主張する一方，文献 [14] は許容度を大きくすれば問題がなくなるとする．では，モデル・キャパシティが小さいと，どうなるだろうか．あるいは，成分ごとに異なる汎化性能を選択的に持たせることができたら，どうなるだろう．

今，入力データが含む性質の異なる成分ごとに汎化性能が異なる DNN モデル M と M' を得ることができたとしよう．ここで，DNN モデル M は，純化成分とゴースト成分の両方を適切に汎化する [9]．一方，M' は，通常の意味での汎化性能が劣り，純化成分に関わる汎化は期待通りであるが，ゴースト成分を適切に汎化できるほどのモデル・キャパシティを持たないとする．M と M' はクリーン・データに対して同等の正確性を示すが，M' は M よりも敵対的な正確性が劣るとする．

図 2 に示すようなアーキテクチャ構成とし，通常汎化版の M と敵対擾乱減衰版 M' を配置する．2 つの予測推論プログラムの出力を比較する予測齟齬の方法で，敵対データの実行時検知を実現する可能性があるだろう．

訓練データが純化成分とゴースト成分に分解できると仮定すると，訓練過程に関して，次のような興味深い議論ができよう．純化成分は当該正解タグの再現に必要不可欠な情報に関わるので，過学習の恐れがあるものの，高い正確性につながる．一方，ゴースト成分は正解タグの再現に補助的な役割を果たす情報からなる．同時に，未知データの予測に有用な働きを示し，その結果，高いロバスト性を達成する．従来からの機械学習の方法は，過学習を避けると共に，ロバスト性を高めることを大きな目的としていた．

*2 現在の DNN モデルに比べて指数関数的に大きなスケールを必要とする．

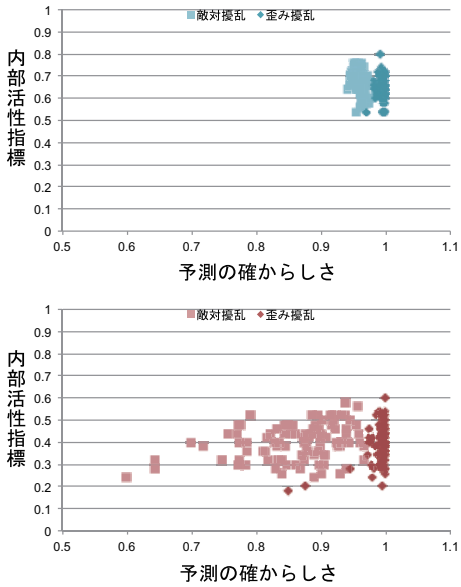

図 3 予測の確からしさと内部活性指標

そこで，多様な訓練データを準備すると同時に，学習方式を工夫することで，敢えて，ゴースト成分から得られる情報を活用していた，と考えてみよう．

求めたい M' はロバスト性に劣る DNN モデルである．このようなモデルを積極的に利用することは，従来の機械学習の研究では想定していなかった．如何にしてロバスト性の高い DNN モデルを得るかが，機械学習研究の主流である．ただし，純化成分に対しては，M' は M と同等の汎化性能を示すようにしたい．果たして，このような M と M' を系統的に得ることができるのだろうか．実験によって確認する．

3.2 異なる汎化性能の実験

最初に，実験の手順を整理する．具体的には，手書き数字分類学習タスクの MNIST データセットを用いる．

(1) 学習モデルを定めて，標準的な学習機構（バックプロパゲーション，確率勾配法，等）[6][8] を実現する 2 つのプログラム $\mathcal{L}_f$ と $\mathcal{L}'_f$ によって訓練した DNN モデルを各々 M と M' とする．ただし，$\mathcal{L}'_f$ は，ロバスト性に劣り，適切な汎化性能を示さない DNN モデル M' を導く [12]．

(2) 実験データセットを第 2.2 節に示した L-BFGS によるセマンティック・ノイズ生成法を用いて作成する．MNIST の試験データセットから抽出したデータの集まり TS_0 を元データとして，敵対擾乱からなるデータの集まり AS と歪みデータの集まり DS を生成する．この時，DNN モデル M の情報を利用するホワイトボックス攻撃の方法を用いた．その結果として，$\langle \vec{x}^{(j)}, t^{(j)} \rangle \in TS_0$ に対して，$\langle \vec{x}_A^{*(j)}, t^{(k)} \rangle \in AS$ $(t^{(j)} \neq t^{(k)})$ と $\langle \vec{x}_D^{*(j)}, t^{(j)} \rangle \in DS$ を得る．

(3) M と M' が，AS と DS に対して示す予測結果の確からしさを測定する．

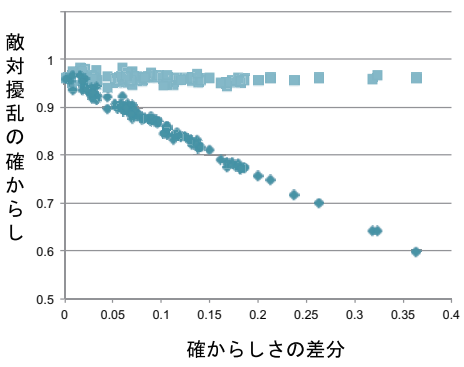

図 4 予測確からしさの違い

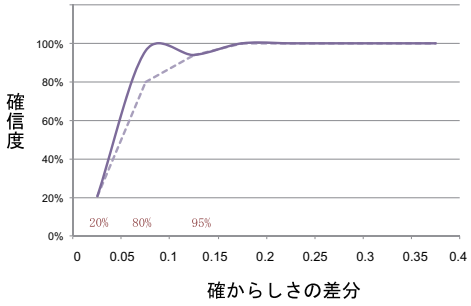

図 5 分離の基準

(4) 測定結果から，M と M' が期待通りに，AS と DS に対して異なる予測確からしさを示すかを確認し，次いで，敵対データを検出する閾値を決定する．

3.3 汎化実験の結果

次に，実験結果のグラフを図 3 から図 5 に示す．

3.3.1 予測確からしさ

図 3 は横軸に予測の確からしさ，縦軸に内部活性指標 [13] を表す．また上段のグラフは M による結果，下段は M' である．各々，AS の要素である敵対擾乱を薄い四角で，DS の要素である歪みノイズを濃い菱形で表した．グラフから以下を読みとることができる．

(図 3-1) 歪みノイズの予測確からしさは，M と M' で差がない．一方，敵対擾乱について，M は M' よりも良い予測確からしさを示す．また，M' は確からしさのバラツキが大きい．

(図 3-2) M は M' よりも大きな内部活性指標値をとる．これは文献 [13] の結果を再現し，$\mathcal{L}'_f$ はロバスト性に劣る DNN モデル M' を導出したことを示唆する．

3.3.2 差分

今，敵対擾乱データを $\vec{x}^{*(n)}$ とする時，縦軸は予測確からしさを表し，濃い菱形は M による予測結果（$Prob(M, \vec{x}^{*(n)}, c^*)$），薄い四角は M' による結果（$Prob(M', \vec{x}^{*(n)}, c^*)$）を表す．横軸は，その差 d である．グラフから以下を読みとることができる．

(図 4-1) 薄い四角の M は，予測確からしさのバラツキが小さく，すべての測定値が 0.95 付近に分布する．AS に対して，敵対的な正確性が概ね良い値を示す．

(図 4-2) 濃い菱形の M' は，敵対的な正確性について，値のバラツキが大きく，0.6 から 0.95 程度まで分布する．

(図 4-3) M' による予測確からしさが劣る敵対データは M による結果との差分が大きく，単調減少の相関を示す．また，M' が良い確からしさの敵対データは差分が小さく，予測結果から M と区別つかない．

3.3.3 分離の基準

図 3 と図 4 から，M と M' は，敵対擾乱と歪みデータに対して異なる汎化性能を示す，と推測できる．また，敵対擾乱に対する M' の予測確からしさは M に比べて劣る．つまり，差分を用いると予測齟齬による敵対データ検知ができる．しかし，図 4 から，予測確からしさが良い時，M と M' は共に同様な振舞いを示し，差分 d から敵対擾乱と歪みノイズを区別することが難しい．d について閾値を決めることにする．

図 5 に，横軸に確からしさの差分 d をとり，縦軸に M の場合のデータ数 $N(d)$ と M' の場合のデータ数 $N'((d)$ として，$N'(d)/(N(d)+N'(d))$ の比グラフを示した．図 3 と合わせると，大きい d の領域では敵対擾乱が大部分を占めることがわかる．

そこで，縦軸の数値指標を「敵対擾乱であることの確信度」と読み替える．ただし，図 5 は実験から得た測定結果であることから，滑らかな曲線で近似して，保守的に確信度とする．たとえば，差分 d が $0.05-0.1$ の時，確信度 80%で敵対擾乱であると推定する．また，d>0.15 では，確信度 95%以上で敵対擾乱として良いだろう．

このような推定の方法が妥当であるかは，別途，新たな評価用データセットによって調べる．

4. 実行時分離の評価

4.1 実験の手順

図 5 に関連して導入した分離の信頼度が有用かどうかの検討を目的として新たな実験を行う．また，クリーン・データを敵対データと判定する誤検出，ならびに敵対データがすり抜ける見落としが生じる確率を測定する．第 3.2 節を参照して，実験の手順を整理する．

(1) 第 3.2 節 (1) で得た DNN モデル M と M' を用いる．

(2) 第 3.2 節 (2) と同様に，M の情報を用いるホワイトボックス攻撃の方法で，第 2.2 節の方法で評価用データセット ES を作成する．ここでは，制約条件 $\Psi(\vec{x}, \vec{x}_S)$ を工夫して，図 1 に示したような多種多様な敵対データならびに敵対性を持たない通常データ（歪みノイズ）を生成する．生成したデータセット ES の構成を調べる．

(3) M による予測確からしさが α 以上のデータの集まりを ES から選び，$ES_{0.9}$，$ES_{0.93}$，$ES_{0.95}$ とする．ここで，α は M による予測確からしさとする．M による予測確か

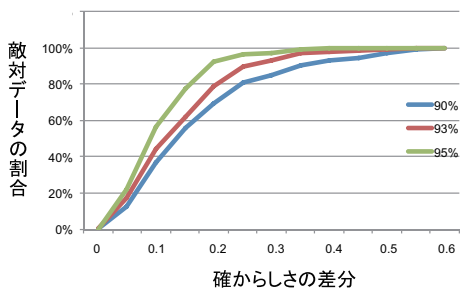

図 6 評価用データセット

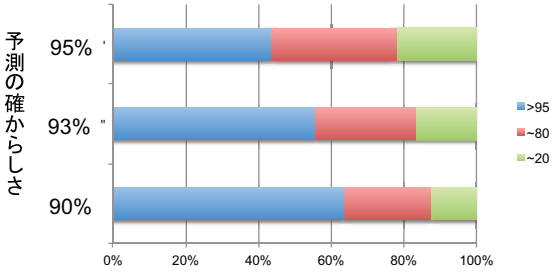

図 7 分離結果

らしさが小さい入力データは，予測結果を利用しない等をアプリケーション固有の判断で行えば良い．本実験の評価対象から外すこととした．

(4) ES_α に対して，M と M' の差分に基づく予測齟齬によって，図 5 の基準から敵対データを検知する．

4.2 分離性能の評価

図 6 と図 7 に ES_α に対する実験結果を示す．

4.2.1 評価用データセットの構成

図 6 は評価用データセット $ES_{0.9}$，$ES_{0.93}$，$ES_{0.95}$ の内容を示す．構成の方法から，大きさが $ES_{0.95}\subseteq ES_{0.93}\subseteq ES_{0.9}\subseteq ES$ となる．図 6 の横軸は予測確からしさの差分 d であり，縦軸は差分が 0 から d までの範囲に入る敵対データの割合を表す．3 種類の ES_α に対してプロットした．グラフから以下がわかる．

(図 6-1) M による予測確からしさ α が大きくなると，小さい差分値 d までに含まれる敵対データの割合が大きくなる．$ES_{0.9}$ は大きな d に対応する敵対データ数が多いことから，小さい d までに含まれる割合が相対的に小さい．

(図 6-2) $ES_{0.95}$ では d<0.2 程度で敵対データの 90%以上を含み，また d<0.1 でも 40%を含む．つまり，$ES_{0.95}$ が含む敵対データの多くが差分の小さい領域に多く存在する．

4.2.2 敵対データの分離性能

図 7 は敵対データの分離結果を示す．棒グラフの各々は，敵対データと結論付けた際の大まかな確信度の比率を表す．最左の青（95%以上），真中の赤（80%程度），最右の緑（20%程度），の順に確信度が下がる．

(図 7-1) $ES_{0.9}$ に着目する．その 60%強を 95%以上の確信度で敵対データと分離したことを示す．つまり，この評価用データセットの場合，敵対であると判定したデータの 60%強が確信度 95%以上で分類できたことになる．これを確信度を 80%程度に条件を緩めると，90%弱の敵対データを検知できている．

(図 7-2) α の値が大きくなるに伴って（棒グラフの下から上へ），分離確信度の高い割合が減少する．これは，(図 6-2) の観察と整合し，最も大きな α 値の $ES_{0.95}$ では d の小さい領域の敵対データの割合が大きいことを示している．つまり，本稿の予測齟齬の方法で分離することが困難な敵対データの割合が大きい．

(図 7-3) 通常データを敵対データと判断する見かけの検知が生じる．$ES_{0.9}$ について，確信度 95 に相当する%d>0.15（図 5）では見かけの検知を 10%以下とすることができる一方で，敵対データ検知率は 60%強（図 7）である．仮に，d>0.1 の領域を敵対データと判定すると，見落としと見かけの検知が均衡し，共に 15%程度となる．実用的とないえない．

5. 考察

5.1 評価用データセットの特徴

図 8 から図 10 は，ノルムの観点から実験データセットの特徴を示すグラフである．以下，$\langle \vec{x}^{(j)}, t^{(j)}\rangle \in TS_0$ に対して，$\langle \vec{x}_A^{*(j)}, t^{(k)}\rangle \in AS$ $(t^{(j)}\neq t^{(k)})$ と $\langle \vec{x}_D^{*(j)}, t^{(j)}\rangle \in DS$ である．

図 8 はセマンティック・ノイズの大きさをノルムで計算した結果を表す．横軸は DS の要素 $\vec{x}_D^{*(j)}$ に対して，縦軸は AS の要素 $\vec{x}_A^{*(j)}$ に対して，3 つのノルム L_0，L_2，L_∞，の計算値をプロットした．ただし，各々のノルムが 0 から 1 になるように正規化した．用いるノルムによって異なる分布を示し，また，敵対ロバスト半径の値が大きく異なることがわかる．

図 9 は L_2 による計算値を取り出したもので，横軸は図 8 と同様に正規化した L_2 値を示し，縦軸は元データ $\vec{x}^{(j)}$ の M による予測確からしさを用いた．薄い緑の四角が表す歪み擾乱の正規化 L_2 値は $0-0.04$ に分布する．濃い緑の菱形が表す敵対擾乱の値は >0.04 に分布する．グラフから，歪み擾乱と敵対擾乱とが，$0.038-0.041$ で分離できることがわかり，この値が敵対ロバスト半径 δ_A に相当すると考えられる．また，予測確からしさが小さく 0.6 以下のデータが存在し，これが，δ_A 近傍に位置する．M による予測確からしさが悪い元データ $\vec{x}^{(j)}$ は，敵対ロバスト半径を曖昧にする．

図 10 は，データセット $ES_{0.9}$ のデータの中から，イメージ上特徴的な 4 つのグループをプロットした．横軸は図 9 と同じであり，縦軸は M による予測確からしさである．

図 9 に関連した推測した敵対ロバスト半径 δ_A が有効かを調べる．歪み擾乱と敵対擾乱は，ほぼ直線上に並び，δ_A

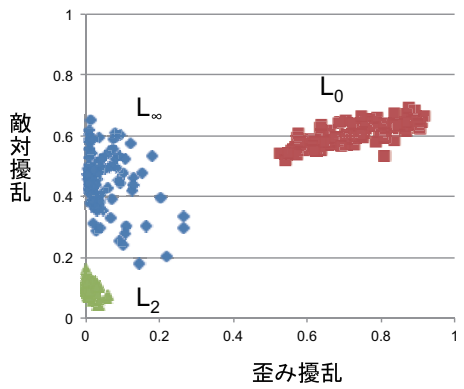

図 8 ノルムからみた特徴

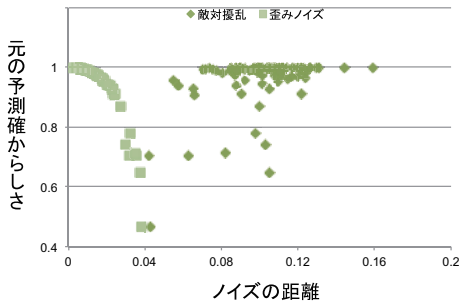

図 9 敵対ロバスト性半径

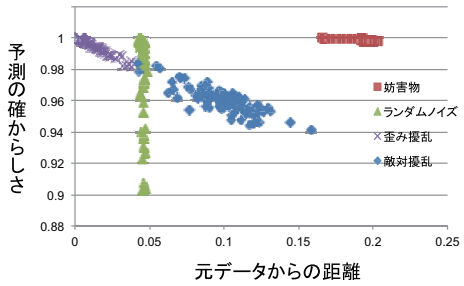

図 10 距離と予測確からしさ

付近で切り替わる．一方，横軸上 0.05 付近の縦長の領域はランダムノイズを付加した通常データ（図 1 の (c) 右上下）右上の横長の領域は妨害物を有する通常データ（図 1 の (c) 左上）である．通常データなので予測結果は正解タグと一致し，また，予測確からしさが良い値であることは，DNN モデル M のロバスト性が優れていることを示す．さらに，これらのデータは図 9 から得られた δ_A よりも元データから離れた領域に存在するが，敵対性を示さない．すなわち，イメージ空間でのノルム（上記では L_2 ノルム）からの議論は，敵対ロバスト半径に対して適切でない．L_2 ノルム値が敵対データの場合よりも大きいにも関わらず，基準となる元データと同じ分類結果となる．

5.2 関連研究との比較

本稿で用いた敵対データ検知の予測齟齬に基づく方法は，さまざまなバリエーションを考えることができる．

文献 [21] は入力データ変換と予測齟齬を組み合わせた敵対データ検知法を提案した．データ $\vec{a}$ を DNN モデル M への入力する．ベクター成分の値を圧縮するような入力変換 $\mathcal{T}^{sq}_{(k)}$ $(k = 1, 2)$ を導入し，$\mathcal{T}^{sq}_{(k)}(\vec{a})$ を得る．3 つの出力値 $(Prob(M, \vec{a}, c)$，$Prob(M, \mathcal{T}^{sq}_{(1)}(\vec{a}), c)$，$Prob(M, \mathcal{T}^{sq}_{(2)}(\vec{a}), c))$ に対して予測結果を比較し，齟齬がある時，入力データ $\vec{a}$ が敵対性を持つと推定する．3 つの予測推論系で用いる DNN モデル M は同様とし，入力データ変換 $\mathcal{T}^{sq}_{(k)}$ の方法を工夫する．

また，通常の訓練による M と敵対訓練によって得た M' にデータ $\vec{a}$ を入力し予測齟齬の方法を利用することも考えられる．敵対訓練は，既知の敵対データが想定通りの正解タグを再現するように DNN モデル M' を訓練する（第 2.3 節）．入力データ $\vec{a}$ が敵対性を持てば，M と M' が異なる分類結果を導く．敵対訓練の特徴から，予め想定した敵対データを検知する法である．

本稿では，特定のセマンティック・ノイズに対して選択的に汎化が劣るモデルを M' として用いた点が異なる．一般に，機械学習の研究は，汎化性能に優れた DNN モデルを得る技術開発を目的としていたことから，本稿のように「意図的に汎化が劣る DNN モデルを構築する」という発想は出てこない．また，本稿の手法は入力データ変換を組み合わせることが可能である．たとえば，$\mathcal{T}^{sq}_{(k)}$ の選び方が予測齟齬による敵対データ検知の性能に影響することが報告されている [21]．

5.3 方法論の妥当性

本稿は，データセットを用いる実験的な手法によって，敵対データを実行時検知する方法を論じた．統計的な分析を行うもので，(1) 予測モデル構築用データセット（第 3 節），(2) 予測モデル評価用データセット（第 4 節）を自動生成し実験で用いた．しかし，一般的な知見を得られるかについては、慎重を期すべきだろう．意図しないで、評価用データセットに対して良い結果が得られるようなチューニング作業を行っているかもしれない．この方法論は，DNN が代表する機械学習の技術と共通する．しかし，一般に用いられていることを理由として，その方法論が妥当と論じることは適切でない．

第 4 節で生成した評価用データセット ES と異なる方法で生成したデータセット $ES^{\dagger}$ を用いて予測モデル評価を行った．敵対データが学習モデル間での転用可能性を持つ [7] ことを利用し，第 2.2 節の自動生成法で用いる学習モデルを $\vec{y^{\dagger}}$ に変更した．この $\vec{y^{\dagger}}$ によって，MNIST データセットを用いて得た訓練済み学習パラメータの集まり $W^{\dagger *}$ とし，A_{λ} で用いる，第 4 節と同様に，条件 $\Psi(\vec{x}, \vec{x}_S)$ をを変えて多種多様なセマンティック・ノイズありデータを生成して，データセット $ES^{\dagger}$ を得た．また，同様に $ES^{\dagger}_{\alpha}$ を抽出する．

生成した敵対データの数は第 4 節の ES_α の場合に比べて大幅に少ない．以下，最も大きな集合である $ES^\dagger_{0.9}$ について考える．統計的な信頼度は低いが，第 4 節と同様に，80%が差分 > 0.1 の領域で検知できた．しかし，通常データの比率が高く，見かけの検知が大きくなった．

差分が大きな通常データを調べると，図 1(c) の左上，枠形状の妨害物によるものが多い．実験で用いた DNN モデル M' は，これらのセマンティック・ノイズを適切に汎化しなかったといえる．枠形状の妨害物を示すセマンティック・ノイズは，DNN モデル M' にとって，敵対的なセマンティック・ノイズと区別できない．一方，図 1(a) や歪みノイズに対しては，差分が微小（< 0.02）であり，DNN モデル M と同等の予測確からしさを示した．このような微小な敵対擾乱に対して，M' は想定通り汎化性能が劣ることを示唆する．

最後に，実適用に際しては，見かけの検知を如何に減らすか，という問題が残る．予測齟齬から得られる情報だけでは確実な検知が難しい．上記の $ES^\dagger$ による実験では，枠状の妨害物を形造るセマンティック・ノイズを，敵対データ検出という観点で適切に取り扱えなかった．しかし，目視からわかるイメージ空間での特徴が異なることから，入力データ変換を併用すれば検知能力が改善すると思われる．このような組合せ手法が重要になる．

6. おわりに

訓練データが純化成分とゴースト成分からなるという仮説をもとに，成分ごとに汎化性能の異なる DNN モデルを利用し予測齟齬の方法による敵対データ検知実験を行った．予測齟齬が生じない場合があり，見落としおよび誤検出をなくすことが難しい．実適用に際しては，当該システムが対象とするデータ特徴を考慮して見かけの検知を減らす方法との組合せが必須だろう．

謝辞　本研究の一部は JSPS 科研費 JP18H03224 の助成および NEDO の委託業務[*3]によって実施した．実験に協力して下さった今井克則氏（アイティスミス）に感謝する．

参考文献

[1] N. Carlini and D. Wagner: Towards Evaluating the Robustness of Neural Networks, In *Proc. IEEE SSP 2017*, and also arXiv:1608.04644, 2016.
[2] N. Carlini and D. Wagner: Adversarial Examples are not Easily Detected: Bypassing Ten Detection Methods, In *Proc. AISec 2017*, and also arXiv:1705.07263, 2017.
[3] E. Eykholt, I. Evtimov, E. Fermandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song: Robust Physical-World Attacks on Deep Learning Models, In *Proc. CVPR 2018*, and also arXiv:1707.08945, 2017.
[4] R. Feinman, R.R. Curtin, S. Shintre, and A.B. Gardner: Detecting Adversarial Samples from Artifacts, arXiv:1703.00410, 2017.
[5] I.J. Goodfellow, J. Shelens, and C. Szegedy : Explaining and Harnessing Adversarial Examples, arXiv:1412.6572, 2014.
[6] I. Goodfellow, Y. Bengio, and A. Courville : *Deep Learning*, The MIT Press 2016.
[7] I. Goodfellow, P. McDaniel, and N. Papernot: Making Machine Learning Robust Against Adversarial Inputs, *Comm.ACM*, vol.61, no.7, pp.56-66, 2018.
[8] S. Haykin : *Neural Networks and Learning Machines (3ed.)*, Pearson India 2016.
[9] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry: Adversarial Examples Are Not Bugs, They Are Features, arXiv:1905.02175, 2019.
[10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu: Towards Deep Learning Models Resistant to Adversarial Attacks, arXiv:1706.06083, 2019.
[11] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frosard: DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks, arXiv:1511.04599, 2016.
[12] S. Nakajima: Distortion and Faults in Machine Learning Software, *Post-Proc. SOFL+MSVL 2019* (to appear), also as arXiv:1911.11596, 2019.
[13] 中島震：訓練済み機械学習モデル歪みの定量指標，電子情報通信学会ソフトウェアサイエンス研究会報告，2020.
[14] P. Nakkiran, Adversarial Robustness May Be at Odds with Simplicity, arXiv:1901.00532, 2019.
[15] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman: SoK: Towards the Science of Security and Privacy in Machine Learning, arXiv:1611.03814, 2016.
[16] K. Pei, Y. Cao, J. Yang, and S. Jana : DeepXplore: Automated Whitebox Testing of Deep Learning Systems, In *Proc. 26th SOSP*, pp.1-18, 2017.
[17] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus: Intriguing properties of neural networks, In *Proc. ICLR 2014*, 2014.
[18] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry: Robustness May Be at Odds with Accuracy, arXiv:1805.12152, 2018.
[19] D. Warde-Farley and I. Goodfellow: Adversarial Perturbations of Deep Neural Networks, in *Perturbation, Optimization and Statistics*, The MIT Press 2016.
[20] W. Wei, L. Liu, M. Loper, S.Truex, L.Yu, M.E. Gursoy, and Y. Wu: Adversarial Examples in Deep Learning: Characterization and Divergence, arXiv:1807.0005, 2018.
[21] W. Xu: Feature Squeezing: Detecting Adversarial Examlpes in Deep Neural Networks, In *Proc.NDSS 2018*, and also arXiv:1704.01155, 2018.

[*3] NII へは再委託