Vol.2020-SE-206 No.4
2020/11/27

ニューラルネットワーク・ソフトウェアの頑健性検査

中島 震[1,a)]

概要：ニューラルネットワーク・ソフトウェアの品質を正解率や適合率から議論することが多い．訓練・学習時に想定しない入力データに対しても，それなりの結果を返すことが期待される．そこで，開発段階に適切な頑健性レベルを持つかを検査しておきたい．本稿では，頑健性をニューラルネットワーク・ソフトウェアの利用時品質と考える．また，検査容易であって頑健性と相関のある指標を提案する．

キーワード：深層ニューラル・ネットワーク，利用時品質，検査指標，教師あり分類学習

Robustness Checking of Neural Networks Software

Shin Nakajima[1,a)]

Abstract: Quality of neural network software is mostly evaluated by means of accuracy or precision. Additionally, such a software system is expected to respond appropriatelly even when input data are not what anticipated at the time of the training. It needs systematic methods, in the development process, to evaluate the robustness levels that the software system exhibits. This paper studies those robustness issues in regard to neural networks, and proposes new testing indices correlated with the robustness and easy to measure as a property of the use-time software qualities.

Keywords: Deep Neural Networks, Use-time Quality, Test Indices, Supervised Machine Learning Classifier

1. はじめに

深層ニューラル・ネットワーク（DNN）に代表される機械学習の技術が進展すると共に，DNN のソフトウェアとしての品質が関心を集めている [11]．そこで，社会的な要請から，DNN ソフトウェア開発に関わる品質マネジメント・ガイドライン [14] が，産官学の協力によってまとめられた．品質基準と達成目標を示すことで，機械学習ソフトウェア品質について，ステークホルダ間での情報共有，利用者への具体的な説明を可能にする．

一般に，DNN ソフトウェアの機能は訓練・学習時に用いるデータ（訓練データセット）によって決まる．訓練データから得られる機能性能が期待通りであると共に，訓練時考慮しなかった想定外のデータに対しても妥当な機能性能を示すことが望まれる．先のガイドライン [14] は，訓練済み学習モデルの品質として，前者を正確性（Correctness），後者を安定性（Stability）と定義している．分類学習タスクを具体例として考える時，正確性は，想定する入力データに対する正解率や正解に分類する予測の確からしさに関わる．

安定性は，未知のデータに対する予測結果を対象とした性質である．従来のソフトウェア・システムで考えると，機能仕様が想定しない（事前条件を満たさない）例外的なデータが入力された時であっても，深刻な結果を生じないことを指す頑健性（ロバスト性）が対応する．検査の方法は，例外系テスティング，ファジング（ファズ・テスティング），負荷テスティング，と呼ばれている．

本稿は，ガイドラインの安定性に着目し，訓練済み学習モデルが示す頑健性（Robustness）の良し悪しを事前に評価する方法について論じる．検査容易であって，頑健性と相関の強い指標を実験を通して検討する．

以下，第 2 節で頑健性についての従来の考え方を概観す

[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

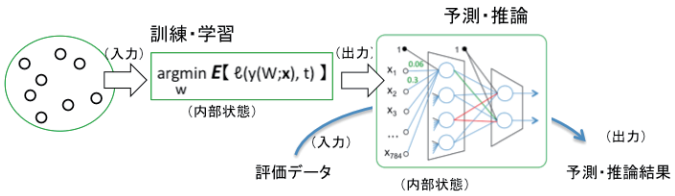


図 1 検査の対象

る．第 3 節で実験によって頑健性予測指標を検討する．第 4 節で提案検査法の適用実験事例を示し，第 5 節で提案法を考察する．

2. 品質からみた頑健性

頑健性に関わる研究は，従来から，統計学や機械学習の分野で活発に行われきた．これを概観する．

2.1 頑健性とは

従来のソフトウェア・システムの頑健性は，機能仕様が規定しない例外的なデータが深刻な結果を生じないことを指す．例外系テスティング，負荷テスティング，の方法が提案されている．例外的なデータを系統的に生成する方法に注目したファジング（ファズ・テスティング）が，オープンなソフトウェアのテスティングで使われている（例えば [11] 第 4 章）．

統計学 [16] は，与えられたデータの集まり（標本）を説明する近似的な関係（統計モデル）を得る方法を提供する．データの母集団分布 ρ を考え，対象の標本が分布 ρ に対して外れ値（Outliers）を含む場合であっても良い近似を得たい．混入した外れ値に対処する方法をロバスト統計 [5] と呼ぶ．事前に外れ値のデータを決定し除去する方法，外れ値の影響を考慮して統計モデルを決定する方法，与えられた標本に対して適切な統計モデルを選択する方法 [2] などがある．つまり，頑健性とは，標本に混入した外れ値の影響が問題にならない統計モデルが示す性質のことである．

機械学習については具体的な議論を行うことから，階層型ニューラル・ネットワークあるいは深層ニューラル・ネットワーク（DNN）に基づく教師あり学習タスクを具体的に考える．DNN[6][7] は，入出力関係が既知の具体的なデータの集まり（訓練データセット）が与えられた時，その入出力を近似的につなぐ関数アルゴリズムを帰納的に求める方法であり，アルゴリズミック・モデリング [3] と呼ばれる．この時，頑健性は，従来のソフトウェアの場合と同様に，訓練データセットに含まれなかったデータが入力された時の予測・推論結果に対する性質である．

機械学習の研究では，頑健性に優れた学習方式・学習アルゴリズムが関心事になる．一般に，得られる訓練済み学習モデルが，訓練データセットに過適合することが問題になる．この過学習を避けることを目的として，さまざまなヒューリスティックス（正則化法）が提案されてきた（例えば [9]）．つまり，訓練データセットに偏りがあると訓練済み学習モデルは頑健性に劣るといえる．

表 1 検査の方法

	検査対象	検査技術
1	訓練データ	統計指標による分析（統計学）
2	訓練・学習プログラム	学習過程の監視（機械学習）
3	予測・推論プログラム	ソフトウェア・テスティング

訓練データセットに偏り（標本選択バイアス）がある場合，期待の機能を示す訓練済み学習モデルを得ることができない．期待通りか否かは，入力データに対する予測・推論結果から判断されることから，入力データが訓練データセットと統計的な特徴が異なる（データ・シフト）時にも不具合が生じる．訓練データセットとシフトの問題は機械学習の主要な研究テーマのひとつである [13]．

次に，頑健性を，大域的な敵対頑健性 [8] を参考に定式化する．基準となるデータ x_c の予測・推論結果 $\mathcal{I}(x_c)$ が既知とする．今，$\|_\|$ を距離とし，微少量 ϵ に対して，$\|\mathcal{I}(x)-\mathcal{I}(x_c)\|\leq\epsilon$ となるような半径 δ（$\|x-x_c\|\leq\delta$）を定義する（定義 [2-1]）．所与の ϵ に対して，δ が期待通りになる時，期待される頑健性が満たされるといえる．関数 $\mathcal{I}(_)$ が連続値をとる時，基準データ x_c から距離の離れたデータ x に対して，予測・推論結果の差が大きくなると考えられる．つまり，その値は滑らかに単調増加する．

分類学習タスクでは，異なる分類結果を導く最大の δ_R を頑健性半径とする．つまり，$\|x-x_c\|\leq\delta_R$ の時 $\mathcal{I}(x)=\mathcal{I}(x_c)$ であり，$\|x-x_c\|>\delta_R$ の時 $\mathcal{I}(x)\neq\mathcal{I}(x_c)$ である．また，x として敵対データを選ぶ時，最大の δ_A を敵対頑健性半径[*1]とする．なお，大域的な敵対頑健性 [8] は，L_∞ ノルムを選んで定義している（$\|x-x_c\|_\infty\leq\delta_A$）．本稿では対象問題に適したノルムを選ぶとする．

2.2 頑健性の検査

定義 [2-1] は，頑健性の特徴を表す一方で，対象の関数 $\mathcal{I}(_)$ あるいは訓練済み学習モデルが持つ頑健性レベルの検査法を与えるものではない．本稿の目的は，頑健性の良し悪しを調べる方法を得ることである．

一般に，DNN ソフトウェアに関して，標本あるいは訓練データセットを対象とする入力空間での検査，訓練・学習過程での検査，予測・推論プログラムを対象とする検査，がある．図 1 に対応して検査方法の概要を表 1 にまとめた．

訓練データを対象とした外れ値に関する分析の方法がある（例えば [1]）．これらの統計的な方法は入力空間で実施する検査に対応する．訓練・学習プログラムを対象とする方法は，学習進行過程の監視のように，機械学習の分野で用いられている検査である．典型的には最適化の対象にな

[*1] 多くの場合，$\delta_A<\delta_R$ である．最近の文献（例えば [4][15]）は敵対頑健性半径を頑健性半径と呼ぶ．

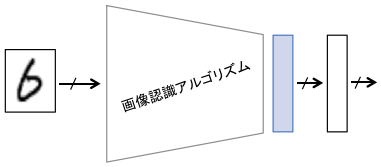


図 2 学習モデル

る誤差関数値の変化に着目する検査がある．訓練データセットへの過適合（過学習）が生じるかを試験データセットに対する正解率によって調べるのは，訓練・学習の状況を知ることが目的なので訓練・学習過程が検査対象である．しかし，直接の検査対象は予測・推論プログラムなので，表 1 では項番 3 に分類できる．この相違は，テスト・オラクルと観測関数の違いによって説明できる [10]．

なお，従来のソフトウェア・テスティング（表 1 項番 3）に関して，多数のテスト・データを用いて統計的な視点から評価する場合，テスト・オラクルに統計的な分析法を用いる方法も有効なことがわかっている [10]．

さて，先の頑健性の定義では $\mathcal{I}(x_c)$ の値に言及していない．一方，ソフトウェア品質を論じる立場では，$\mathcal{I}(x_c)$ が適切な値になること，つまり，正確性が良いことが同時に重要である．まとめると，所与の基準データ x_c に対する $\mathcal{I}(x_c)$ が期待通りであること，検査対象 x が基準データからの距離 δ に応じて期待される ϵ になること，といえる．

本稿では，DNN ソフトウェアの利用時品質として，頑健性のレベルを開発時に検査する．頑健性を考慮した学習方式・学習アルゴリズムを用いたか否かに関わらず，$\mathcal{I}(_)$ の頑健性の良し悪しを調べることである．そこで，品質検査を目的とするソフトウェア・テスティングの一貫として，テスト・データならびにテスト・オラクルの選択法を考える．

3. 検査指標

テスト・オラクルに関わる検査指標を検討する．

3.1 検査の観点

頑健性の良し悪しの事前見積もりに適した指標が何かを検討する．素朴には，先の定義から，予測・推論結果 $\mathcal{I}(_)$（予測の確からしさ）がわかれば良い．一方，このような外部指標は，訓練済み学習モデルへの多様な影響を隠してしまう [11]．事前見積もりには，内部の状態を参照するほうが好ましい可能性がある．そこで，ニューロン・カバレッジ [12] をもとにした内部活性状態を内部指標として導入する．どのような指標が目的に合うかは，後に予備実験によって確認する．以下，画像分類学習タスクを具体例として検討する．

3.1.1 外部指標

まず，外部指標を定義する．予測・推論プログラム $\mathcal{I}(x)$ の実体は，画像認識アルゴリズムを実現する層に加えて，最終段の出力層の活性関数に softmax を持つとする（図 2）．分類カテゴリ数を C とする時，入力データ a に対する $\mathcal{I}(a)$ の出力値は C 次元ベクトル $p[c]$（$c = 1, \ldots, C$）である．$p[c]$ の各成分は確率を表すので，$p[c]\in[0,1]$ である．また，$c^* = \underset{c\in[1,C]}{argmax}\ p[c]$ が入力データ x に付与された正解タグ t と一致する時，予測・推論結果が正解となる．この時，$p[c^*]$ は c^* と推定した予測の確からしさを表す．たとえば，$p[c^*]$ 値が小さい場合であっても，他の $p[c]$ との相対的な比較で最大になれば，正解と判断されることになる．

外部指標は，予測・推論プログラム $\mathcal{I}(_)$ を対象とするソフトウェア・テスティングのテスト・オラクルであるが，単独の特定データ a に対する $\mathcal{I}(a)$ ではなく，評価用データの集まり ES に対して求める．つまり，統計的なソフトウェア・テスティングを考えていることになる．具体的には，正解の頻度から正解率（Accuracy）を，また，予測の確からしさ $p[c^*]$ の平均 $\overline{p}$ を求めることができる．これらは正確性を表す良い指標である．

3.1.2 内部指標

次に，内部指標を定義する．ニューロン・カバレッジ NC は，対象ニューラル・ネットワーク中の着目するニューロン n 全体の個数 N に対する活性ニューロン（Active Neurons）の比率として定義される．入力データ a による信号伝播によって生起されるニューロン n の出力値を σ_n とする．別途，閾値 τ が与えられた時，$\tau<\sigma_n$ となるニューロンの個数 A が決まる．この時，ニューロン・カバレッジ NC を A/N と定義する．$NC\in[0,1]$ が成り立つ．

本稿の内部指標は着目するニューロンを最後から 2 番目の層（Penultimate Layer），つまり，図 2 の網掛けの層を対象とした内部活性指標とする [11]．前段までの画像認識アルゴリズムに対応する層の計算結果が此の層に集約されることが理由である．

3.2 予備実験

3.2.1 実験の対象

予備実験は MNIST データセット[*2]を対象として行う．学習モデルは古典的なニューラル・ネットワークで全接続（fully connected）とし，中間層（図 2 の画像認識アルゴリズム）の活性関数を $ReLU$ とする．内部指標の計算に際して，閾値 τ を 0 とした．

まず，中間層のニューロン数が異なる学習モデルを準備し，2 つの訓練・学習プログラム PC と BI を用いて訓練済み学習モデルを求める．これまでの実験（[11] 等）から，BI によって得た予測・推論プログラム $\mathcal{I}_{BI}$ は PC による $\mathcal{I}_{PC}$ に比べて頑健性が劣ることが確認されている．すなわち，訓練データ x_c に対して同等の結果（予測の確からし

[*2] http://yann.lecun.com/exdb/mnist/

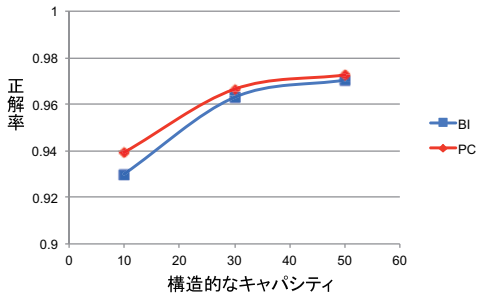


図 3 正解率の変化

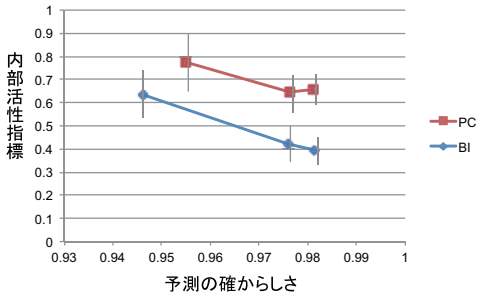


図 4 内部活性指標の変化

さ）を返す場合であっても，x_c にノイズを付加した評価データ x に対して $\mathcal{I}_{BI}(x)$ は $\mathcal{I}_{PC}(x)$ よりも予測の確からしさが小さくなる．$\delta_c = \| x - x_c \|$ が決まる時，予測の確からしさ値を比較して，$\epsilon_{BI}(\delta_c) < \epsilon_{PC}(\delta_c)$ である．

3.2.2 正解率の違い

図 3 は中間層ニューロン数（構造的なキャパシティ）の異なる 3 つの学習モデルを用いて構築した $\mathcal{I}_{BI}$ と $\mathcal{I}_{PC}$ の正解率を比較したグラフである．ここで，評価データの集まり ES として MNIST の試験データセット TS を用いた．グラフから，構造的なキャパシティが大きくなると，共に正解率が向上することがわかる．つまり，中間層のニューロン数を増やすと表現力が向上するが確認できる．一方，キャパシティが小さいと正解率に有意な差が見られるものの，$\mathcal{I}_{BI}$ と $\mathcal{I}_{PC}$ でそれほど大きな違いがない．実際の開発では，十分な構造的キャパシティの学習モデルを用いることから，正解率について，$\mathcal{I}_{BI}$ と $\mathcal{I}_{PC}$ は同等といえる．

3.2.3 内部活性指標の違い

図 4 は縦軸に内部活性指標をとった．予測の確からしさの平均 $\bar{p}$ が良くなると共に，$\mathcal{I}_{BI}$ と $\mathcal{I}_{PC}$ は共に単調減少の傾向を示す．図 3 から正解率や $\bar{p}$ は構造的なキャパシティと正の相関があるので，構造的なキャパシティが大きい時に内部活性指標が減少することは，予測・推論に有効に関わる必須の活性ニューロン数が，構造的なキャパシティの増加に比べて大きく増えないことを示す．逆に，構造的なキャパシティが小さい時に内部活性指標が大きくなることは，少ない構成ニューロンをフルに活用していることを示唆する．つまり，冗長性が小さいといえる．

また，内部活性指標の絶対値を比較すると，$\mathcal{I}_{BI}$ は $\mathcal{I}_{PC}$ よりも小さい．したがって，$\mathcal{I}_{BI}$ では，予測・推論に有効に関わらない構成ニューロンが多い．一方，構造キャパシティが不明な時，この絶対値の情報のみから，$\mathcal{I}_{BI}$ と $\mathcal{I}_{PC}$ を区別することができない．つまり，$\mathcal{I}_{BI}$ が頑健性に劣るとしても，内部指標の絶対値によって頑健性の違いを議論できない．

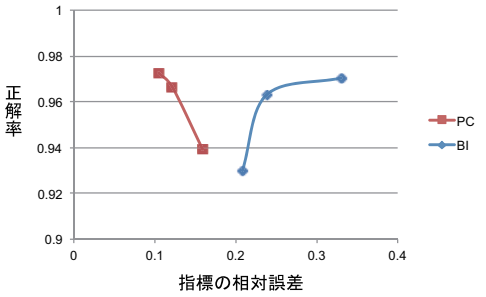


図 5 相対誤差の変化

3.2.4 頑健性の検査指標

図 5 は，横軸に内部活性指標の相対誤差を選んだ．内部活性指標の標本平均を μ，標本分散を σ とする時，相対誤差を σ/μ とした．これによると，$\mathcal{I}_{BI}$ と $\mathcal{I}_{PC}$ の 2 つのグラフは明らかに異なる傾向を示すことがわかる．$\mathcal{I}_{BI}$ は正解率，すなわち，構造キャパシティが大きくなると共に σ/μ が大きくなる．他方，$\mathcal{I}_{PC}$ は σ/μ が小さくなる．

同一の構造キャパシティで両者を比較すると，μ の値は $\mathcal{I}_{BI}$ が $\mathcal{I}_{PC}$ よりも小さい（図 4）．つまり，$\mathcal{I}_{BI}$ は σ が大きいことを示し，評価データ a に対する活性ニューロン数のバラツキが大きいことを示す．ところが，経験的に頑健性が良いとわかっている $\mathcal{I}_{PC}$ は，バラツキが小さく，評価データ a の違いに依らず，多数の構成ニューロンを平等に有効利用していると考えられる．

3.2.5 定性的な考察

頑健性と σ/μ が示すバラツキの関係について定性的に考察する．画像分類のタスクについて，学習モデル（図 2）は入力ベクトル成分（ピクセル）の相関を分析しながら分類に不要な情報を除去するフィルターの役割を果たすと考えて良いだろう．

今，訓練データとの統計的な違いから生じるデータの差をノイズとする時，ノイズを含むデータは入力ベクトル x が分類に寄与しない情報を多く含み，フィルターを通過する情報が少ない．その結果，予測の確からしさ $p[c^*]$ が低下する．頑健性が良い $\mathcal{I}_{PC}$ は，ノイズを持つデータに対して，最終的な分類結果を期待通りに出力するように，適切な水準のフィルターが作動する．その結果，想定外のデータに対する予測推論結果が大きくて低下しない，つまり，安定性に優れていると解釈できるだろう．

一方，頑健性が劣る $\mathcal{I}_{BI}$ はフィルターが適切に働かないことから除去する情報量が大きいと考えられる．ノイズは訓練データを基準として想定外の情報であることから，バ

リエーションが大きく，また，不適切なフィルターが除去する情報のバリエーションが大きい．その結果，入力ベクトルによって活性化されるニューロン数，したがって，内部活性指標のバラツキが大きい．そこで，σ/μ が示すバラツキとして現れる．逆に，評価データに対して，このようなバラツキが大きいと，検査対象は期待する頑健性を持たないと推定できる．

4. 提案の検査法

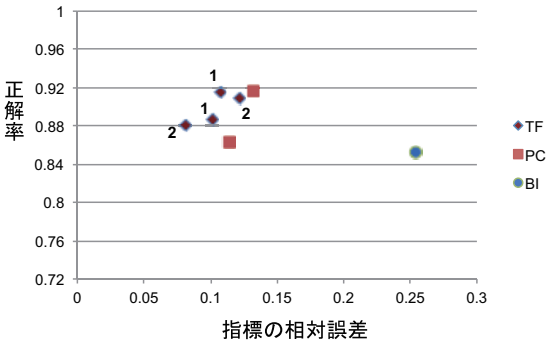
4.1 検査手順と適用例

予備実験（第 3.2 節）の結果から，訓練済み学習モデルが示す頑健性の良し悪しを評価する指標として，内部活性状態の σ/μ を選ぶ．ここで，内部活性状態は，学習モデルの Penultimate 層（図 2）を対象とするニューロン・カバレッジである．

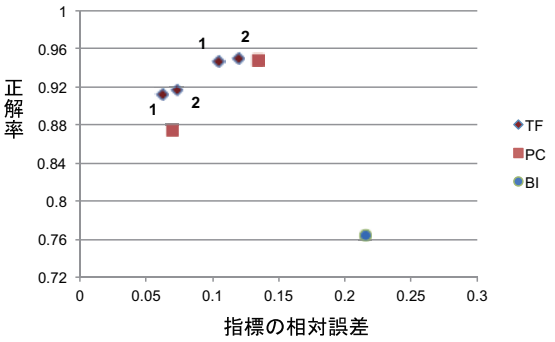
実験では，教師あり分類学習タスクについて，複数の学習データセット，複数の学習モデルを用いて，どのような検査結果になるかを調べる．内部活性指標の相対誤差 σ/μ を頑健性の良さの相対的な比較指標として用いることであり，頑健性の絶対的な評価を目的とするものではない．所与の学習タスクに対して，複数の訓練済み学習モデルを得た時，正確性に加えて，頑健性に優れる対象を選ぶ方法を論じることを目的とする．以下の手順で調べる．

(1) 対象分類学習タスクのデータの訓練データセットを LS，試験データセットを TS とする．
(2) 学習モデル $y(W;_)$ ならびに訓練・学習を実行する最適化の方法を決める．
(3) 訓練・学習プログラムを実行して訓練済み学習モデル $y(W^*;_)$ （$W^* = \underset{W}{argmin}\, \mathcal{E}_{\langle x,t\rangle}[\![\, \ell(y(W;x),t)\,]\!]$）を得る．
(4) $y(W^*;_)$ から構成した予測・推論プログラム $\mathcal{I}(_)$ の返却値が C 次元ベクトル p であって，ベクトル p の第 c 成分 $p[c]$ がクラス c に分類される確からしさを表すとする．
(5) 試験データセット TS の全ての要素 a について，$\mathcal{I}(a)$ を実行し，内部活性状態の σ/μ，予測の確からしさが最大となるクラス c^* （$c^* = \underset{c}{argmax}\, p[c]$）の値 $p[c^*]$ の平均値 $\bar{p}$，ならびに正解率を求める．
(6) 横軸に σ/μ，縦軸に $\bar{p}$ をプロットする．頑健性を重視する時は横軸の小さいものを選ぶ．横軸の値が同等であれば縦軸の大きいものを選べば良い．

実験では，データセットとして，Fashion MNIST[*3] と Kuzushiji MNIST[*4] を用いた．予備実験で用いた MNIST と同じ形式のデータ（28×28 のピクセル）からなり，実験結果の比較が容易なことが理由である．



(a) Fashion MNIST



(b) Kuzushiji MNIST

図 6 実験の結果

学習モデルは CNN[6] の方法を採用し，ネットワーク構造の異なる 2 つ学習モデル（CNN1 と CNN2）を比較した．Fashion MNIST と Kuzushiji MNIST は，MNIST よりも複雑な画像でなので，全結合では妥当な正確性を得ることが難しい．画像認識の性能に優れた CNN を用いる．

また，訓練・学習を実行するフレームワークも 2 つを比較した．ひとつは予備実験で用いたフレームワークで本稿を含む一連の研究 [11] で採用したもの（基本版），もう一方は Google から公開されている Tensorflow/Keras[*5] である．さらに，訓練学習の最適化（更新パラメータ計算法）についても 2 種類（SGD と Adam）を実験した．実験結果を図 6 に示す．

4.2 適用実験の結果

図 6 で，基本版で得た訓練済み学習モデルを四角（■）と丸（●）で示し，Tensorflow 版を菱形（◆）で表した．基本版では，■が予備実験の PC に，●が BI に対応する．予備実験と同様に，BI による σ/μ 値（横軸の値）は PC に比べて大きいことがわかる．つまり，図 2 の画像認識アルゴリズムに対応するネットワークの種類に依らずに予測実験の場合と同様な傾向を示すと言って良い．

Tensorflow 版では 2 つの CNN を比較実験し，図 6 中，1（CNN1）あるいは 2（CNN2）を付した．なお，基本版の

*3 https://github.com/zalandoresearch/fashion-mnist/
*4 http://codh.rois.ac.jp/kmnist/
*5 https://www.tensorflow.org/

ⓒ 2020 Information Processing Society of Japan

PC と BI は共に CNN1 である．また，基本版と Tensorflow 版の 3 種類について各々 2 つの最適化手法の結果を示しているが，すべての組合せについて正解率が良いのは Adam だった．一方，横軸の値は Adam のほうが大きい．本稿の議論の流れからは，相対的にみて，Adam は正確性が良いが，頑健性は若干劣ると解釈することになる．

次に，基本版と Tensorflow 版を比較するので共通する CNN1 の結果に着目する．正確性と頑健性の両方から，Tensorflow 版が平均して妥当な傾向を示すとして良いだろう．Tensorflow が基本版と比較して，良くエンジニアリングされた学習フレームワークということがわかる．

5. 考察

第 2.1 節で概観したように，統計学や機械学習の分野で多数の研究がある．特に頑健性半径を事後的に求める方法について，文献 [17] はモンテカルロ法の重点サンプリングと同様な考え方で，δ を効率良く求める評価用データの生成方法を提案する．訓練済み学習モデルあるいは $\mathcal{I}(_)$ の頑健性の良し悪しを検査するものではない．

文献 [15] は L_∞ ノルムで特徴つけられる敵対データを防御する敵対性学習（Adversarial Learning）の研究であって，正確性と敵対頑健性が両立しないことを論じている．訓練データ数を増やして正確性を向上させようとすると，データが密になり，敵対頑健性半径 δ_A が小さくなるというものである．敵対頑健性を対象として両立できないことを理論的に示した．

図 6 の結果に関連して，Adam と SGD の比較から，相対的にみて，Adam は正確性が良いが，頑健性は若干劣ると解釈することを述べた．DNN では，Adam が正確性に優れていることが良く知られている（[6][7] 等）．一方，頑健性については論じられていない．

6. おわりに

本稿では，従来のソフトウェア・システムに期待される頑健性と同様な観点から，深層ニューラル・ネットワークによる予測・推論プログラムの頑健性を，その開発時に見積もる方法を論じた．絶対的な頑健性を評価する方法ではなく，所与の学習タスクに対して，複数の訓練済み学習モデルを得た時，頑健性に優れると考えられる対象を選ぶ方法を論じることを目的としている．

一般に，DNN は有限という意味で不完全な訓練データから適切な汎化能力を示す入出力関係を近似的に求める方法であり，素朴に考えて，ガイドライン [14] の正確性と安定性は両立しない．開発目的に合わせて，どちらを優先するかを決めなければならず，それ故，開発時に正確性ならびに頑健性を検査しておく必要がある．本稿の議論は，実験による経験的な方法に基づくもので，提案方法の妥当性が一般的に確認されたわけではない．理論面からの議論が必要である．

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[*6]によって実施した．機械学習の頑健性について議論して下さった磯部祥尚氏（産総研），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] C.C. Aggarwal: *Outlier Analysis (2ed.)*, Springer 2017.
[2] 赤池弘次，甘利俊一，北川源四郎，樺島祥介，下平英寿: 赤池情報量規準 AIC，共立出版 2007.
[3] L. Breiman: Statistical Modeling: The Two Cultures, *Statistical Science*, 16(3), pp.199-231, 2001.
[4] N. Carlini and D. Wagner: Towards Evaluating the Robustness of Neural Networks, In *Proc. IEEE SSP 2017*, and also arXiv:1608.04644, 2016.
[5] 藤澤洋徳：ロバスト統計，近代科学社 2017.
[6] I. Goodfellow, Y. Bengio, and A. Courville : *Deep Learning*, The MIT Press 2016.
[7] S. Haykin : *Neural Networks and Learning Machines (3ed.)*, Pearson India 2016.
[8] G. Katz, C. Barrett, D. Dill, K. Julian, and M. Kochenderfer: Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks, In *Proc. CAV 2017*, pp.97-117, 2017, and also arXiv:1702.01135v2, 2017.
[9] G. Montavon, G.B. Orr, and K.-R Mukker (eds): *Neural Networks: Tricks of the Trade (2ed.)*, Springer 2012.
[10] 中島震：統計的な部分オラクルによるテスティング法，日本ソフトウェア科学会第 37 回大会，66-L，2020.
[11] 中島震：ソフトウェア工学から学ぶ機械学習の品質問題，丸善出版 2020.
[12] 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.
[13] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N.D. Lawrence (eds.) : *Dataset Shift in Machine Learning*, The MIT Press 2009.
[14] 産業技術総合研究所 サイバーフィジカルセキュリティ研究センター: 機械学習品質マネジメント・ガイドライン v1.0，CPSEC-TR-2020001，2020.
[15] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry: Robustness May Be at Odds with Accuracy, arXiv:1805.12152, 2018.
[16] 東京大学教養学部統計学教室（編）：統計学入門，東京大学出版 1991.
[17] S. Webb, T. Rainforth, Y.W. Teh, and M.P. Kumar: A Statistical Approach to Assessing Neural Network Robustness, In *Proc. ICLR*, 2019, and also arXiv:1811.07209, 2018.

[*6] NII へは再委託

ⓒ 2020 Information Processing Society of Japan