AIモデルの定量的な頑健性評価手法

Techniques for Quantitative Evaluation of Noise Robustness of AI Models

仲　義行 TSUZUKI Yoshiyuki　大平　英貴 OHIRA Hidetaka　高橋　信太郎 TAKAHASHI Shintaro

近年，最先端のAIモデルをミッションクリティカルな分野である社会インフラや製造分野に適用し，不具合発生の予兆検知や検査の省力化といった運用・保守の高度化や，新たな価値の創出に役立てることが期待されている。このようなAIモデルの高い信頼性を維持するには，AIモデルを搭載したシステム（以下，AIシステムと略記）の品質管理が重要になってくる。

そこで，東芝は，AIモデルの特徴を考慮した品質管理を実施するため，AIシステムに組み込まれるAIモデルの頑健性を定量的に評価する手法を開発した。従来のAIモデルの精度評価に，定量的な頑健性指標を用いた評価を加えて品質管理を行うことで，AIシステムの運用時の信頼性向上に寄与することができる。

The introduction of state-of-the-art artificial intelligence (AI) models to mission-critical systems in the social infrastructure and manufacturing fields not only enhances operation and maintenance services by detecting signs of abnormalities and reducing labor for inspection work, but also makes it possible to offer users new value. These AI systems are now faced with the need for quality management so as to maintain highly reliable AI models.

In order to implement quality management taking the characteristics of each AI model into consideration, Toshiba Corporation has developed techniques to evaluate the noise robustness of AI models incorporated into AI systems through evaluation of their inference performance using quantitative indexes together with conventional techniques. These quantitative evaluation techniques can contribute to the improvement of reliability of AI systems in operation.

1. まえがき

近年，急速に進展してきた機械学習，特にディープラーニングへの注目は一層高まりを見せており，様々な産業分野において，機械学習で生成されたAIモデルを利用した製品やサービスの開発・利用が進められている。

しかし，AIモデルは，従来のソフトウェアのように仕様に基づいて作成された“演繹（えんえき）的”なものではなく，膨大なデータを学習して自律的に答えを導き出す“帰納的”な手法で生成されるものなので，品質を評価・管理することが非常に難しい。そのため，社会インフラや製造分野のようなミッションクリティカルな分野では，AIシステムの実用化は，いまだ限られている。

AIモデルの品質の評価・管理の難しさを解決するため，AIモデルならではの特徴を考慮したテスト手法への関心が高まっている。AIモデルの不確実性の高い振る舞いや，入力データの微小な変動でも出力に大きく影響が出てしまう特徴，従来ソフトウェアのテストと比較してAIモデルに対するテストオラクル（テストの合否を判断する根拠）の生成が困難などの問題に対するテスト技術の研究が行われている[1]。

ここでは，製造分野で製品外観検査への機械学習の適用事例が多い，教師あり学習による画像分類問題を題材として，AIモデルのテスト技術を活用した，“AIモデルの頑健性”に関する品質評価手法について述べる。

2. AIモデルの頑健性

AIシステムの品質を管理するための指針として，AIプロダクト品質保証コンソーシアムが“AIプロダクト品質保証ガイドライン”[2]を発行している。このガイドラインでは，AIモデルに対して考慮すべき品質観点として，AIモデルの入力データに何らかの変化があっても安定して性能を維持できる頑健性を取り上げている。また，品質の高いAIシステムを開発するためには，精度（推論結果の正解率）が高く，かつ頑健性が確保されたAIモデルの生成が重要であると述べている。

具体的には，従来のAIモデルで用いられる精度や汎化性能による評価に加え，“ノイズに対して頑健か”という観点でもモデルを評価することが挙げられている。製品外観検査を例にすると，照明状況の変化や撮影カメラの感度などが要因で画像に対してノイズが入ることが考えられるため，実用ではノイズに対して頑健であることが重要である。

そこで，AIモデルの頑健性を評価するために，AIモデル

のノイズ耐性を定量的な指標とする評価手法を開発した。

3. ノイズ耐性の定量指標化

教師あり学習を用いた画像分類問題では，ラベル付き画像を学習することでAIモデルを生成する。このAIモデルに画像を入力すると，画像に対するラベルが推論される。このAIモデルに入力する画像にノイズを加えた場合，ノイズを大きくすると一般に精度は下がる（**図1**）。

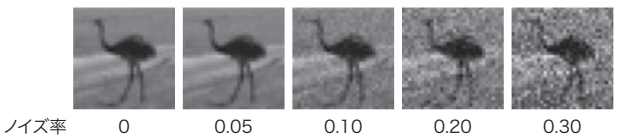
ノイズに対して頑健なAIモデルは，ノイズを加えても推論結果が変わらずに精度を保てる。そこで，推論結果が変わらないノイズの大きさを定量指標化して，ノイズ耐性として頑健性の評価を行う。開発手法では，人が知覚できる偶発的ノイズと敵対的摂動に対し，それぞれで定量指標化した。

敵対的摂動とは，人には知覚できないが，AIモデルが持つ脆弱（ぜいじゃく）性を狙って意図的に生成された微小ノイズのことである。敵対的摂動を付与することで，AIモデルが誤推論するデータを敵対的サンプル[4]と呼び，AIシステムの脆弱性評価に利用されている。

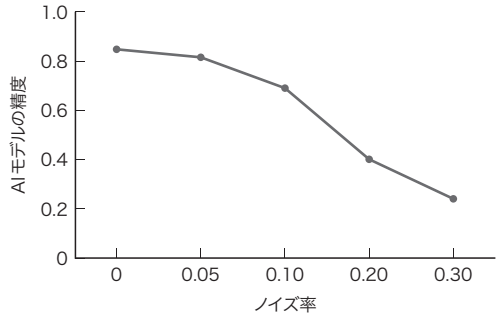
ここで，偶発的ノイズは利用環境などの変化に対して，敵対的摂動はAIモデルのノイズに対する脆弱性に対して，それぞれ頑健性を評価することを想定している。

3.1 偶発的ノイズへの耐性

偶発的ノイズとして，実運用で通常想定される目視可能なノイズを想定する。具体的な例としては，図1で取り上げた画像に対するホワイトノイズが挙げられる。



(a) ノイズを加えた画像の例



(b) ノイズの大きさに対するAIモデルの精度

図1. ノイズによるAIモデルの精度低下の例

画像（CIFAR-10[3]を使用）の各画素に加わるホワイトノイズ（0 ～ 127の乱数にノイズ率を乗じたもの）が大きくなると，画像分類の推論精度が下がる。

Reduction of accuracy rate of image classification due to increase in white noise

開発手法では，Randomized Smoothing[5]（以下，RSと略記）を用いて推論結果が変わらないノイズの大きさを算出し，その結果をPSNR（Peak Signal to Noise Ratio）で定量指標化した。

RSとは，AIモデルの推論結果が変化するノイズの理論的最小値を，ノイズを加えた際に出力される推論結果（画像分類問題であれば，どのラベルが出力されるか）の期待値を用いて算出する手法である。開発手法では，偶発的ノイズによって変化する推論結果の期待値をRSに適用し，推論結果が50 %の確率で変化する偶発的ノイズの大きさを算出可能にした。算出したノイズの大きさは，期待値が最も高い推論結果が2番目に期待値が高い推論結果に変化する最小値である。この最小値より小さい偶発的ノイズであれば，期待値が最も高い推論結果は変わらないことが保証されるため，このノイズの値を基に偶発的ノイズへの耐性が評価できる。

PSNRは，信号が取り得る最大パワーに対するノイズ比率を示すもので，非可逆な画像圧縮での画質劣化指標としても広く用いられている。PSNRは，ノイズがゼロで無限大，ノイズが大きいほど小さい値となる。人間の主観画質とは必ずしも一致しないが，おおむね40 dB以下になると，劣化が知覚されるようになる[6]。推論結果が変化する偶発的ノイズの最小値をPSNRで表現すれば，ノイズ耐性の目標値を定量的に分かりやすく設定できるようになる。

3.2 敵対的摂動への耐性

敵対的摂動は，人が知覚できないほど微小なのでその大きさを実感しにくく，偶発的ノイズのようにノイズ量の大きさによる妥当な基準値を設けることが難しい。そのため，敵対的摂動に対しては，大きさを人がイメージしやすい，ほかの指標が必要であると考えた。

そこで，開発手法では，敵対的摂動を検知するよう学習した検知器を用いて，敵対的摂動を加えたデータ（以下，敵対的データと略記）を含むデータセットに対し，検知器が敵対的摂動を検知する割合（以下，検知率と略記）を測定する。検知率は，言わば“AIモデルにとってのノイズの見分けやすさ”であり，これをノイズの大きさの指標とすることで，より人に分かりやすい目標値を設定できるようにした。例えば，検知率が高い摂動は，AIモデルにとって見分けやすい大きな摂動となる。

4. ノイズ耐性の評価手法

品質の高いAIシステムを実現するために，利用者が求めるAIモデルの精度に目標値を設定するのと同様に，ノイズ

に対する頑健性にも目標値を設定して評価することが重要である。開発手法では，偶発的ノイズと敵対的摂動のそれぞれで，ノイズの大きさの定量指標に対し目標値を設定して評価を行う。ここでは，それぞれの定量指標に目標値を設定して，ノイズ耐性が評価できるかを検証した。検証用の画像データには，CIFAR-10を用いた。

4.1　偶発的ノイズへの耐性評価

偶発的ノイズに対して，RSで算出したノイズ最小値をPSNR化して定量指標とし，これに目標値を設定してノイズ耐性が評価可能かを検証した。

評価では，モデルA，モデルB，モデルCの三つのAIモデルについて，ノイズ耐性を比較した。モデルAは，無加工の学習データを用いて生成した。モデルBは，モデルAの学習データに，ホワイトノイズを加えた学習データを付加して生成した。モデルCは，ホワイトノイズを大きくしてモデルBと同様に生成した。学習条件は，3モデルとも同一とした。

3モデルについて，テストデータに対してRSで算出したノイズ最小値をPSNRで指標化し，ノイズ耐性を測定したグラフを**図2**に示す。グラフは，横軸がノイズをPSNRで指標化した値，縦軸がノイズの影響を受けない割合（PSNR値の大きさのノイズを加えても推論結果が変わらないデータ数の割合）としてプロットした。ここでは，推論結果の正解／不正解ではなく，ノイズによって推論結果が変わる／変わらないで，ノイズ耐性を評価している。

この場合，小さいPSNRでも推論結果が変わらないデータが多いAIモデルほど，ノイズ耐性が高いことになる。図2では，モデルCのノイズ耐性が最も高いことが分かる。

また，人が知覚できるノイズの大きさに該当するPSNR値を参照し，例えば，“70％以上のデータが40 dB以上のノイズに耐性があることを目標値にする”など，目標値を設定して評価することが可能となる。

4.2　敵対的摂動への耐性評価

敵対的摂動に対する検知率に目標値を設定して，ノイズ耐性が評価可能かを検証した。

評価では，モデルD，モデルEの二つのAIモデルについて，ノイズ耐性を比較した。モデルDは，無加工の学習データを用いて生成した。モデルEは，モデルDの学習データに学習データから生成した敵対的データを加えて，同じ学習条件で学習し，よりノイズ耐性を向上させたモデルである。敵対的データの生成手法には，FGM（Fast Gradient Method）[7]を用いた。

それぞれのモデルの検知器を生成して，敵対的データ（テストデータから生成）と，無加工のテストデータとに対する検知率を，敵対的摂動の大きさを変えながら測定した。測定結果を**図3**に示す。グラフは，横軸が敵対的データに対する検知器の検知率，縦軸が敵対的データに対するAIモデルの精度としてプロットした。

図3のepsは，FGMで使用する入力パラメーターの一つで，生成する摂動の大きさに乗じる値として，大きさを調整するために用いる。そのため，epsを大きくすると摂動が大きくなる。図3の結果から，敵対的摂動の大きさと検知率との相関が分かり，検知率を摂動の大きさを表す指標として使用可能なことが確認できた。

摂動の大きさと検知率との相関から，検知率が高いノイズに対しても精度が高いAIモデルが，ノイズ耐性が高いモデルになる。図3では，モデルEがモデルDよりもノイズ耐性が高いと相対評価できることが分かる。

摂動の大きさを検知率で示すことで，例えば，“検知率が

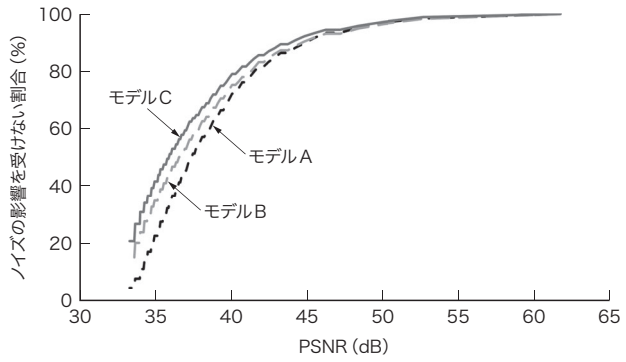


図2.　PSNR化したノイズ最小値で指標化したノイズ耐性評価

テストデータに対するノイズ最小値とノイズの影響を受けない割合の関係をモデルごとにグラフ化することで，ノイズ耐性を評価する。

Noise robustness evaluated using peak signal-to-noise ratio (PSNR) at smallest value of noise as index

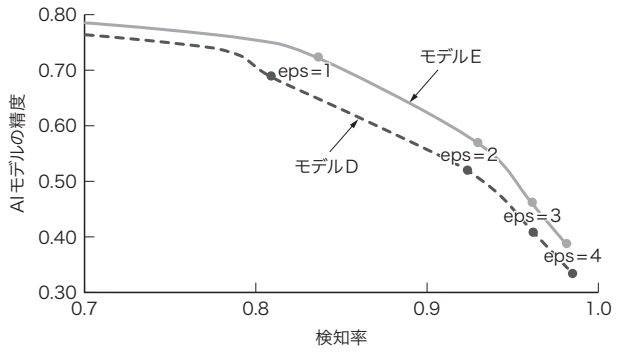


図3.　敵対的摂動の検知率で指標化したノイズ耐性評価

テストデータに対する検知率とAIモデルの精度の関係をモデルごとにグラフ化することで，ノイズ耐性を評価する。

Noise robustness evaluated using adversarial perturbation detection rate as index

0.8以下となる大きさの敵対的摂動に対しては，モデルの精度が70％以上となること”など，定量的かつ人がイメージしやすい目標値を設定して，各々のモデルを評価することが可能となる。

4.3 ノイズ耐性によるAIモデルの頑健性評価

最後に，ノイズ耐性の定量指標値とAIモデルの頑健性の評価方法についてまとめる。

生成したAIモデルの頑健性を，偶発的ノイズと敵対的摂動の観点で評価する。前者は，期待値が最も高い推論結果が変わらないノイズ最小値を各テストデータに対してRSで算出し，ノイズを加えても推論結果が変わらないデータ数を，ノイズの大きさを変えながら測定することで評価する。後者は，テストデータに敵対的摂動を加えたときの摂動の検知率と推論精度との関係を，摂動の大きさを変えながら測定することで評価する。

定量指標化したノイズ耐性は，人が直観的に判断できるものであり，対象とする画像や分類目的に応じ，目標値を設定してAIモデルの頑健性が評価できる。

5. あとがき

教師あり学習を用いた画像分類問題を事例として，ノイズ耐性を用いてAIモデルの頑健性を定量的に評価する手法の開発・検証を行った。

今後は，評価手法をツール化し，AIシステムの開発プロセスへの適用を進めるとともに，AIシステムの信頼性をより高めるために，評価手法を拡充して品質の管理方法の整備を進めていく。

文　献

⑴ Zhang, J. M. et al. Machine Learning Testing: Survey, Landscapes and Horizons. IEEE Transactions on Software Engineering, 2020, arXiv:1906.10742, <https://arxiv.org/pdf/1906.10742.pdf>, (accessed 2021-01-29).

⑵ AIプロダクト品質保証コンソーシアム編．AIプロダクト品質保証ガイドライン 2020.08版．2020，266p．<http://www.qa4ai.jp/QA4AI.Guideline.202008.pdf>,（参照 2021-01-12）.

⑶ Krizhevsky, A. "The CIFAR-10 dataset". The CIFAR-10 and CIFAR-100 datasets. <https://www.cs.toronto.edu/~kriz/cifar.html>, (accessed 2021-01-12).

⑷ Goodfellow, I. J. et al. "Explaining and Harnessing Adversarial Examples". 3rd International Conference on Learning Representations (ICLR 2015). San Diego, CA, 2015-05, Computational and Biological Learning Society. 2015, arXiv:1412.6572. <https://arxiv.org/pdf/1412.6572.pdf>, (accessed 2021-01-12).

⑸ Cohen, J. et al. "Certified adversarial robustness via randomized smoothing". 36th International Conference on Machine Learning (ICML 2019). Long Beach, CA, 2019-06, International Machine Learning Society. 2019, p.2323–2356. arXiv:1902.02918. <https://arxiv.org/pdf/1902.02918.pdf>, (accessed 2021-01-12).

⑹ 小箱雅彦．標準化への道 標準化委員会報告3 電子文書の画像圧縮ガイドライン．月刊IM．2011，**50**，5，p.21–24.

⑺ Dong, Y. et al. "Boosting Adversarial Attacks with Momentum". Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018). Salt Lake City, UT, 2018-06, IEEE. 2018, p.9185–9193. arXiv:1710.06081. <https://arxiv.org/pdf/1710.06081.pdf>, (accessed 2021-01-29).

仲　　義行　TSUZUKI Yoshiyuki
研究開発センター
知能化システム技術センター AI応用推進部
AI Application Dept.

大平　英貴　OHIRA Hidetaka
研究開発センター
知能化システム技術センター AI応用推進部
AI Application Dept.

高橋　信太郎　TAKAHASHI Shintaro
研究開発センター
知能化システム研究所 システムAIラボラトリー
System AI Lab.