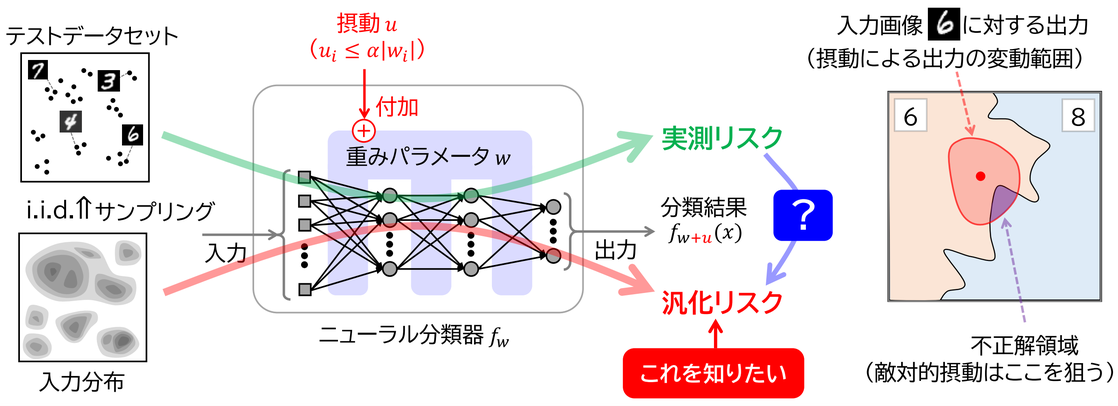
敵対的摂動に対するニューラル分類器の確率的安全性保証

リスク有データ：不正解領域の割合が**許容閾値**を超えるデータ
実測リスク：テストデータ数に対するリスク有データ数の割合
汎化リスク：任意のデータセットに対するリスクの期待値

研究の目的

敵対的摂動*1に対するニューラル分類器*2の確率的安全性保証

- 目標：**実測リスク**から**汎化リスク**上界を見積る**計算方法**の開発[1]



*1 不正解になるように付加する摂動（無作為摂動ではない）　*2 ニューラル分類器：分類器として訓練したニューラルネットワーク

実験

代表的なニューラル分類器の汎化リスク上界（信頼度90%以上）見積り

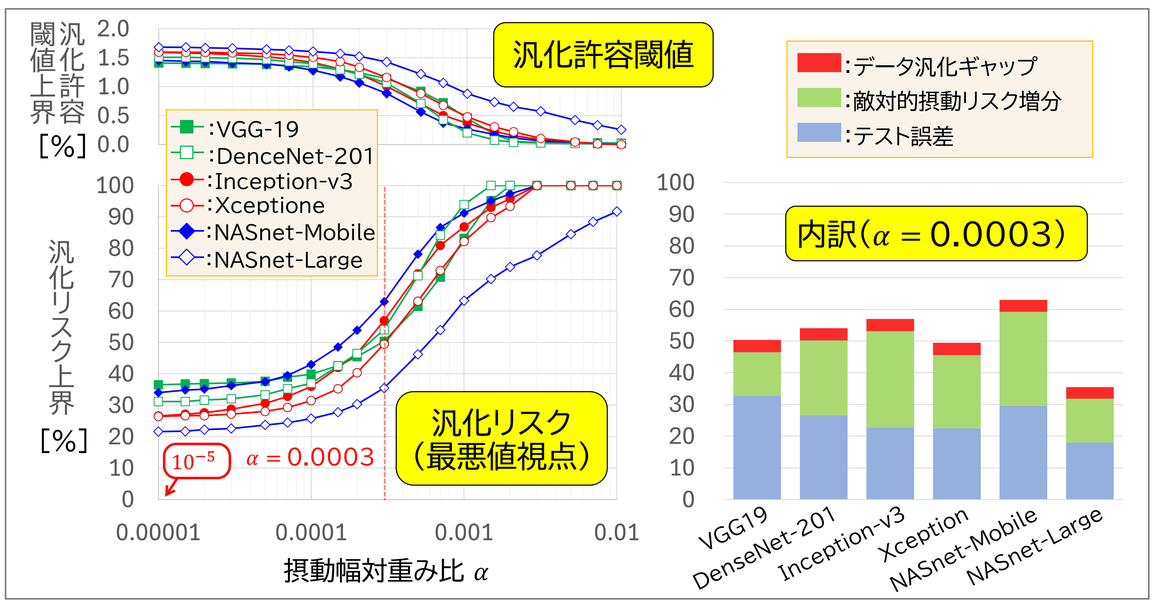
- 分類器：Inception-v3, NASnet-Largeなど（層数：数十～数百）
- テストデータセット（サイズ：1,000）：ImageNet（画像：256 × 256画素）
- **敵対的摂動探索**：I-FGSM（反復勾配法）
- 敵対的摂動を検出できないデータについては …
  - ✓ 確率的に存在しないことを保証（**許容閾値** 2%以下）
  - ✓ 保証のための**無作為摂動テストの摂動サンプル数** m は次式により得られる

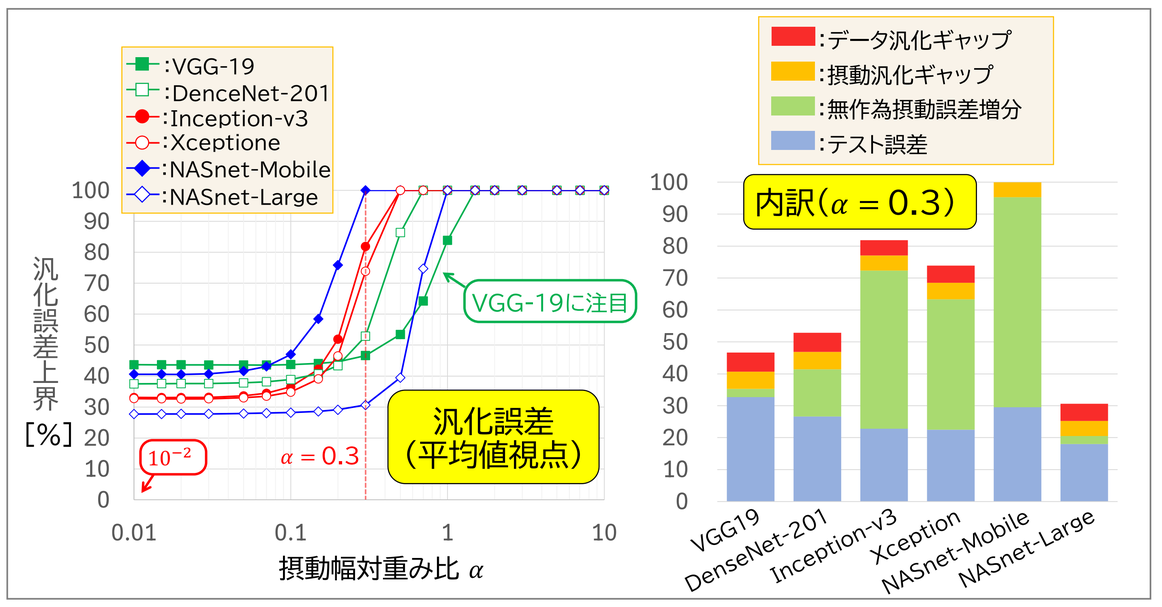
$$m = \left\lceil \log_{(1-\theta^*)}\left(\frac{\delta}{2n_0}\right) \right\rceil$$

（θ^*：許容閾値，δ：不信頼度，n_0：不検出データ数）

計算例．$\theta^* = 0.02,\ \delta = 1 - 0.9,\ n_0 = 800$ならば、$m = \lceil 479.16 \ldots \rceil = 480$

結果　摂動幅対重み比に対する汎化リスク上界（信頼度90%以上）の見積結果（比較：無作為摂動付加汎化誤差（不正解率の期待値）上界）





考察

実測リスクではなく汎化リスクを見積もる理由

利用者視点で意味が明確な値（信頼度や許容閾値など）で説明できる（**データも任意**）

・例．摂動幅対重み比0.0003の場合のニューラル分類器NASnet-Largeのリスクの計算結果の例

『テストデータセット（サイズ1,000）に対する**実測リスク**は31.8%』

『任意のデータセットに対する**汎化リスク**（汎化許容閾値1.4%以下）は35.7%以下（信頼度90%以上）』

（実測リスクの疑問：なぜ1,000？ 他のデータに対する評価は？ 不検出データのリスクは？等）

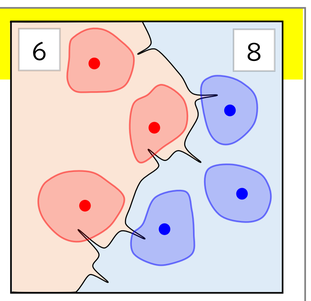
汎化リスク（最悪値視点）と汎化誤差（平均値視点）の違い

汎化リスクと汎化誤差は**異なる観点からニューラル分類器を評価**できる

・**汎化誤差**：自然なノイズに対する耐性の評価に有効

・**汎化リスク**：意図的な敵対的摂動（攻撃）に対する耐性の評価に有効

（上の実験結果よりVGG-19では右図のような微小リスクの存在が多いと考えられる）



敵対的摂動探索と無作為摂動テストを併用する理由

無作為摂動テストのみでは敵対的摂動検出に膨大なサンプルが必要*3

・**探索によって効率よく**敵対的摂動を検出可能

・**無作為摂動テストによって不正解率の上界**を見積可能

妥当な許容閾値の設定（信頼度等も同様）

許容閾値の設定は**計算コストとのトレードオフ**

・許容閾値を下げるには無作為摂動サンプル数を増やす必要がある

・参考：本実験のXceptionの計算時間は1設定あたり約100分

- 産総研スパコンABCI 3.0, tf_HC（16コアCPU），GPU無効

既存技術との違い：実用性と確率的保証の両立

敵対的摂動探索と無作為摂動テストと許容閾値の**適切な組合せ**

・汎化リスクの関連研究：理論的解析[3]、確率的見積法[4]

*3 理論的には無作為摂動のみで不正解になる敵対的摂動を探索した方が簡潔であるが、そのためには膨大な無作為摂動サンプル（例えば、10^{10}以上）が必要

結論

・未見のデータセット/摂動も含めて汎化リスクの上界を確率的に保証

- 機械学習品質マネジメントガイドライン[5]の安定性評価*4に適用可能
- 必要なデータ数や摂動数を信頼度や許容閾値によって説明可能
- 意図的な敵対的摂動（攻撃）に対する耐性の評価に有効

・WP-GEB-Estimator：汎化リスク見積ツール

- 本実験で使用したソースプログラムをウェブサイトで公開中[2]

謝辞　この成果の一部はNEDOの委託業務（JPNP20006）の結果得られたものです

*4 安定性：データセット以外の未見の入力に対しても安定した推論が行われること

参考文献

[1] 磯部, 最悪重み摂動付加ニューラル分類器の汎化誤差上界の見積法, JSAI2024
[2] WP-GEB-Estimator, 2024. https://staff.aist.go.jp/y-isobe/wp-geb-estimator/
[3] Y. Tsai, et al., Formalizing Generalization and Adversarial Robustness of Neural Networks to Weight Perturbations, NeurIPS 2021.
[4] T. Zhang, et al., PRoA: A Probabilistic Robustness Assessment against Functional Perturbations, ECML-PKDD 2022.
[5] 機械学習品質マネジメントガイドライン（第4版）2023 https://www.digiarc.aist.go.jp/publication/aiqm/guideline-rev4.html

磯部 祥尚
ソフトウェアアナリティクス研究チーム
Contact : y-isobe@aist.go.jp

産総研
ともに挑む。つぎを創る。