大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

トップエスイー ソフトウェア開発実践演習

生成モデルを用いたニューラルネットワーク検証における検証領域近似手法の改良

富士通株式会社　　横山晴樹

近似を用いたNN検証の問題点

- 「入力に摂動を与えてもNN が誤認識しないか」等の性質を検証する技術では、スケーラビリティのため検証する領域を近似する
- 性質検証において、近似領域が決定境界を跨いでいると、真偽が曖昧となる

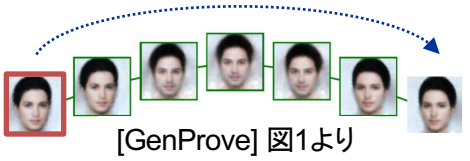
手法の改良

少数のサンプルに対するNNの出力を用いて決定境界からの距離を推定し、近似領域の大きさを制御する

➔ 決定境界を跨ぐ近似領域の削減

近似を用いたNN検証

生成モデルを用いた画像の摂動



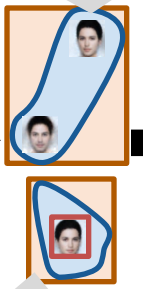
[GenProve] 図1より

性質検証
顔の向きを徐々に反転させたとき同一人物と判定できるか？

入力画像

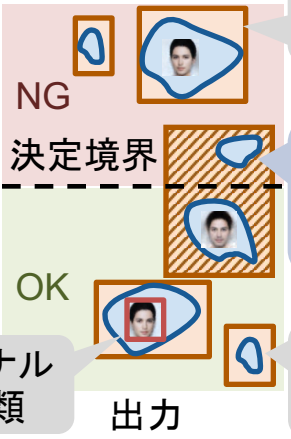
生成モデル

摂動で得られる画像の集合○



近似領域 □

分類モデル



オリジナルの分類

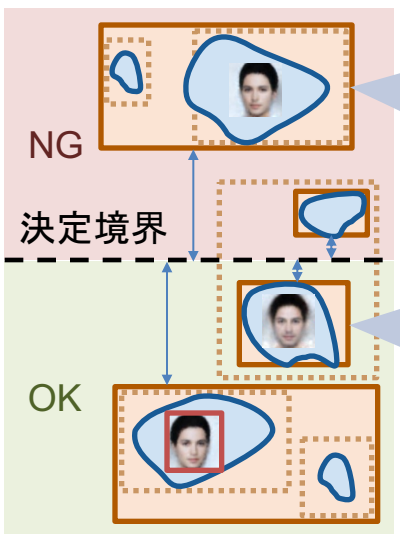
近似領域全体がNG
➔ 元の集合でもNG

近似領域が決定境界をまたぐ
➔ **検証結果が曖昧**

近似領域全体がOK
➔ 元の集合でもOK

改良点

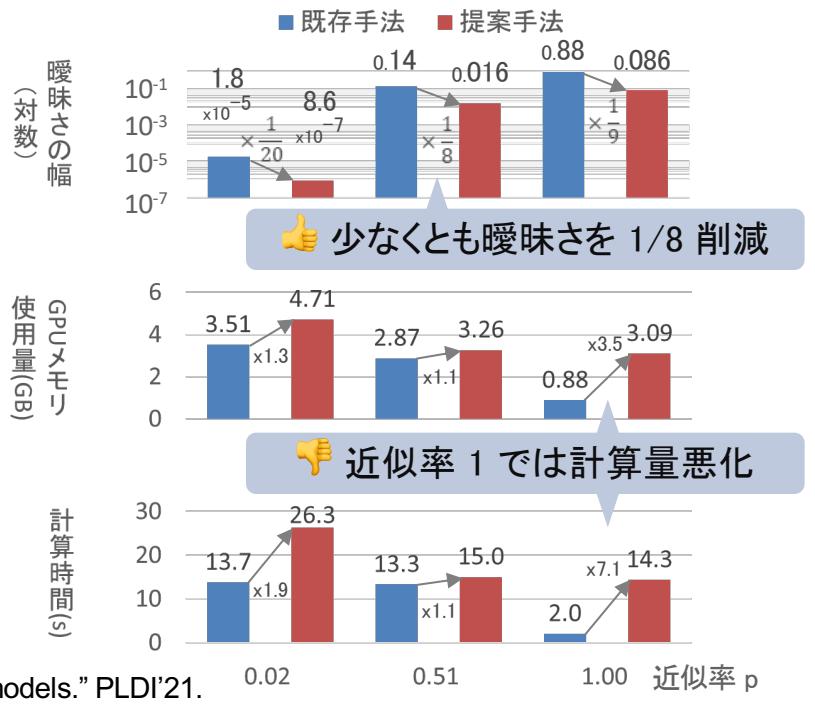
少数サンプルに対し決定境界からの距離を測定
➔ **距離に応じた近似戦略**



決定境界から遠い
➔ 積極的に近似
➔ スケーラビリティ向上

決定境界に近い
➔ 近似を控える
➔ 検証の正確さ向上

評価



👍 少なくとも曖昧さを 1/8 削減

👎 近似率 1 では計算量悪化

[GenProve] Mirman, Matthew, et al. “Robustness certification with generative models.” PLDI’21.