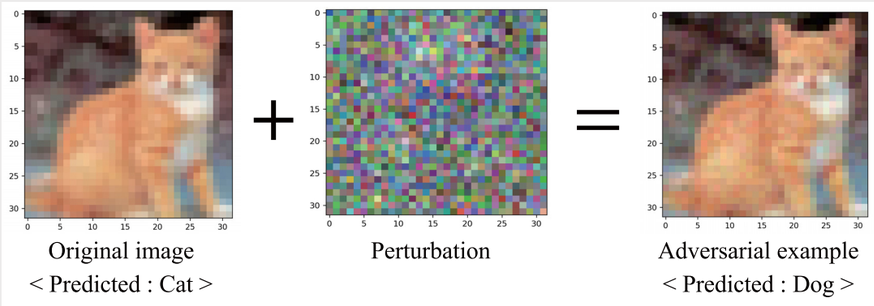
「第36回 独創性を拓く 先端技術大賞」　主催：産経新聞社

特　別　賞

「AI・深層学習への敵対的サンプル攻撃に対する新たな防御手法の提案」

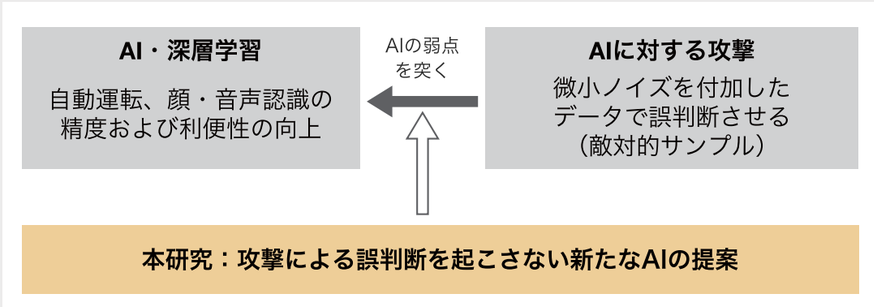
中央大学大学院　理工学研究科　電気・情報系専攻　博士課程後期課程2年　田崎 元

1. 背景



- 人工知能（AI）や深層学習は、悪意あるデータにより欺かれ、誤った判断を起こす危険性が指摘されている。
- 顔認証への他人なりすましや自動運転の誤認識、医療画像処理システムの誤判断など、その被害は甚大である。
- 誤判断は微小なノイズを利用して作られる不正なデータ（敵対的サンプル）で引き起こされ、ヒトの目では本来のデータとの違いに気づけずに、AIが悪用されてしまう。（例：右端のネコの画像はイヌと判断されてしまう）

2. 本研究の目的



課題

- 人工知能の構造や仕組みは複雑であり、これまでは、現象を解明できず、対策の確立に至らなかった。
- 従来手法は敵対的サンプルの誤分類を十分に防げないだけではなく、本来の識別精度を低下させてしまう手法が多い。

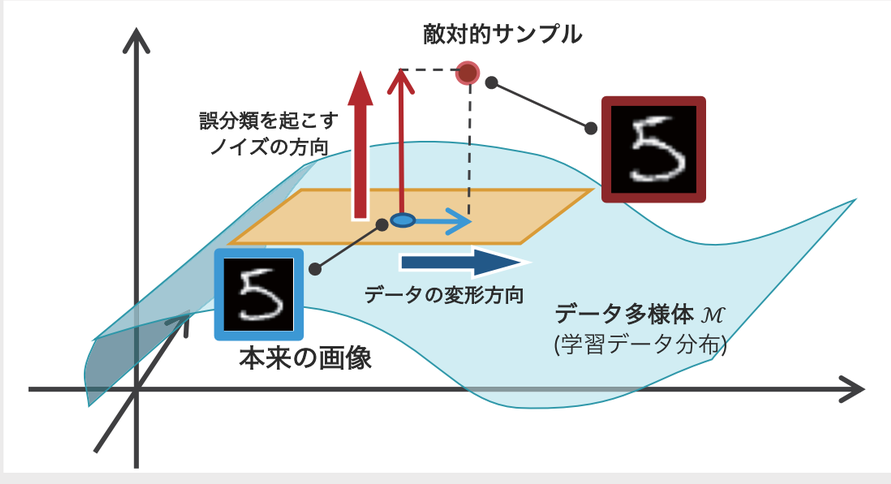
目的

AIが本来持つ識別精度を低下させず、高い精度で敵対的サンプルによる誤分類を防ぐ防御手法を提案する。

3. 提案手法：敵対的サンプルの発生メカニズムと防御手法

敵対的サンプルの発生メカニズム

- AIの学習データは、複雑なデータ分布（多様体構造）を持っており、満遍なく分布するのではなく、空洞や偏りが生じており学習できない領域が多数存在する。
- データが存在しない領域をAIは認識することができず、ランダムに補間してしまうことで、敵対的サンプルによる誤分類が引き起こされる要因となる。
- 誤分類を起こすノイズは学習不十分な領域の方向に作られ、データの変形（例：手書き文字画像の字体や顔画像の表情変化）を示す学習データの分布に直交する。
- この構造を解明したことで敵対的サンプルを正しく分類可能な防御を実現することができる。



敵対的サンプル攻撃に対する防御手法

- 本手法は、AIの判断基準は学習データに依存して決まるため、学習に用いたデータの分布（多様体）構造を解析して、誤分類を起こすノイズ成分を抽出・除去することで、正しい識別を実現する。
- 入力データごとに、データの変形成分（青矢印）と誤分類を起こすノイズ成分（赤矢印）に分解し、データ成分のみを抽出することで、入力が敵対的サンプルであってもノイズ成分を除去することができて正しく識別することができる。さらに、この操作はノイズ成分にだけ作用するため、AI本来の識別精度を低下させることなく導入可能である。

4. 実験

３種類の敵対的サンプル生成手法に防御手法を適用し、正しく分類できた割合（Accuracy）で評価を行った。

- いずれも９割以上の割合で誤分類を防ぐことができ、高い水準で誤分類を防ぐことができた。
- 最も強力な攻撃と呼ばれるC＆W attackに対する防御成功率が最も高く、95.6%の割合で正しく分類できた。
- ノイズのない本来のデータに対するAIの分類精度は、データ成分の抽出が有効にはたらき、精度を低下させることなく、わずかに向上する結果が得られた。

敵対的サンプルに対する防御成功率

生成手法	Accuracy
FGSM ($\varepsilon = 0.05$)	94.5 %
FGSM ($\varepsilon = 0.10$)	91.5 %
PGD	94.4 %
C&W attack	96.5 %

AIの分類精度

	適用前	適用後
Accuracy	98.00 %	98.36 %

5. 結論

- 敵対的サンプルによる誤分類とAIが学習するデータの構造は、これまで明らかにされておらず、今回は多様体理論という現代数学の道具を駆使して、誤分類の原因を突き止めた。さらに、敵対的サンプルの発生メカニズムに基づく防御手法を提案し、 AI本来の識別性能を劣化させることなく、高い水準で防御可能であることを示した。
- 本手法は AI研究の理論・実践だけではなく、今後AIに求められる安全性への貢献が期待できる。