データ多様体の埋め込み幾何学に基づいた敵対的サンプルの検知手法

Detecting Adversarial Examples Based on Embedding Geometry of Data Manifolds

久重広樹・ネットワーク分科会・中央大学大学院

1.背景

・**AI**が画像認識分野などにおいて**セキュリティ**にまで応用されている。

・敵対的サンプルと呼ばれるAIの**脆弱性**が発見された。

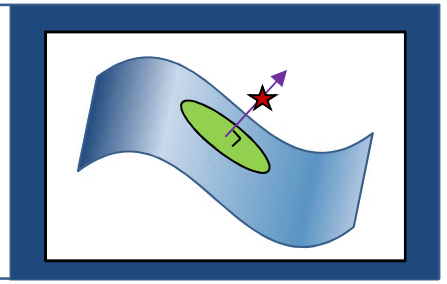
・敵対的サンプルの**発生メカニズム**に基づいた対策手法はなかった。

2.発生メカニズムに基づいた提案手法

・**敵対的サンプル**は正常データが成す**多様体の外**に存在する。(発生メカニズム)

・入力データを中心としたデータ多様体の**接空間**を推定し、**角度**を算出する。

・接空間と入力データ間の角度が大きければ敵対的サンプルと判断する。



3.結果と今後の予定

・本検知手法は、敵対的サンプルと正常データの分布の分離に成功した。

・敵対的サンプルの種類によって検知精度が低下するため、さらなる調査を行う。