データ多様体構造に基づいた敵対的サンプルの生成手法の提案

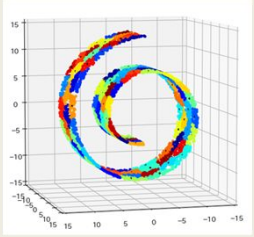
Creating Adversarial Examples Based on Data Manifold

森田 匡博・暗号分科会・中央大学大学院

研究概要

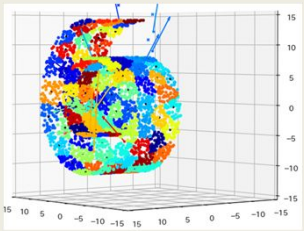
IoT機器や自動運転技術の発展で深層学習の活用機会が増えているが、脆弱性が指摘されているため、過去の攻撃を包括した手法を提案することで防御の構築に役立てる。
深層学習に用いるデータ多様体から、敵対的サンプルを生成する
学習に用いるデータの多様体は空間に充満しておらず、部分多様体を構成する
余次元方向に移動することで分類境界をまたぎ誤認識を起こさせる(以下スイスロールの例)

K-meansで多様体の分割



PCAを適用し寄与率に基づいて次元の算出

余次元(直交方向)への移動



研究結果

MNISTを用いて攻撃を生成
従来の攻撃手法では
発見できなかった
攻撃の存在可能性

実際の攻撃

今後の方針

- 既存手法で生成された攻撃との比較
- 射影空間上での直交方向の算出