データ多様体の接空間推定と深層学習の敵対的サンプルの対策

Tangent Space Extraction of Data Manifolds and Countermeasures Against Adversarial Examples in Deep Learning

宮坂優吾・ネットワーク分科会・中央大学

Abstract - In recent years, the applications of machine learning have expanded, achieving remarkable results in various fields such as image processing, speech processing, and autonomous driving. However, the neural networks at the core of these systems possess a vulnerability known as adversarial examples. Adversarial examples are generated by adding a small perturbation, called an adversarial perturbation, to a normal sample, which can cause machine learning models to misclassify the input without being noticeable to humans. Although various countermeasures against adversarial examples have been proposed, none have provided a complete solution. Tasaki et al. proposed a defense method based on the mechanism of adversarial example generation. While this method has shown promising results on the handwritten digit dataset MNIST, it has not been successful on low-density datasets such as the object color dataset CIFAR-10. In this study, we attempt to improve Tasaki et al.'s defense method by utilizing the intermediate layers of the model as a data manifold and increasing the dataset density through data augmentation to enhance the accuracy of estimating the tangent space of the data manifold.

1.研究背景

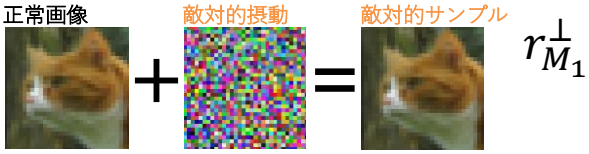
深層学習の脆弱性

近年、深層学習は幅広い分野で発展を遂げている
深層学習の根幹技術であるニューラルネットには敵対的サンプルと呼ばれる脆弱性が存在する

敵対的サンプルとは？

敵対的サンプルはわずかな摂動を加えただけで、人間には変化がわからないがモデルが誤認識するデータ である。
これにより信頼性・安全性に大きな影響を及ぼす。

正常画像　敵対的摂動　敵対的サンプル

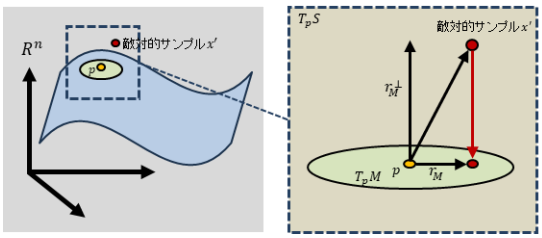


猫の画像に摂動を加えることでモデルは馬と判断した

2.先行研究

データ多様体に基づいた敵対的サンプルの対策

実世界データは高次元空間R^nに埋め込まれた低次元多様体Mとしてモデル化できる。
この多様体は局所的に接空間T_pMによって線形近似が可能。
敵対的サンプルは接空間T_pMに直交する補空間$T_p^{\perp}M$ に存在する



先行研究の課題

入力点に近い正常データを利用するため密度の低いデータセットを利用すると接空間の推定がうまくいかない（ほかのラベルの多様体が混じってしまう）

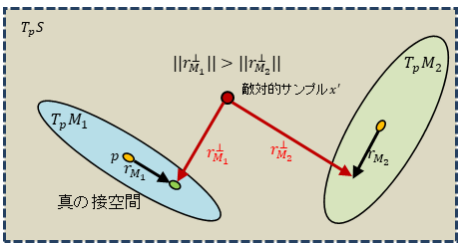
3.提案手法

2つの方法を利用しデータ密度を上げることで先行研究の課題を解決する。

①モデルの中間層の特徴の抽出
　中間層の出力をデータ多様体として扱う
②データ拡張
　近傍点に対してデータ拡張を行う

防御アルゴリズム

1. モデルの中間層の特徴の抽出①
2. 一次元高い空間への埋め込む
3. 近傍の作成
4. 近傍内で同ラベル近傍の作成
5. 同ラベル近傍のデータ拡張②
6. 各ラベルごとの接空の推定
7. 各ラベルごとの直行補空間成分を抽出
8. 直交補空間成分が最小のものを真の接空間と選定
9. 入力点を接空間へ埋め込む



4.実験結果

3 種類の敵対的サンプルに対する提案手法の防御性能を,CNN モデルを用いて評価した.防御性能の評価には,攻撃データ 10,000 枚(攻撃の成否を考慮しない)を用い,防御後のデータに対する分類精度を比較することで検証を行った.

攻撃手法	防御前	先行研究	提案手法
FGSM	51.12%	34.71%	51.38%
PGD	41.53%	34.68%	42.60%
C&W	17.19%	34.74%	52.81%

すべての攻撃手法に対して先行研究の手法よりも高い分類精度を達成した.特に,現在最も強力と云われる C&W Attackに対しては,防御適用後の分類精度が大幅に向上している.一方で,FGSM および PGD に関しては,防御前後で分類精度の向上が限定的であった.
一方,PGD と FGSM に関しては,敵対的サンプルの生成時にモデルの損失関数の勾配に対して直接最適化するため,モデルの中間層の出力においても正常データと類似した特徴を持つ可能性がある.その結果,提案手法による識別が難しくなったと考えられる.これに対し,C&W Attack はモデルの損失関数を直接使用するのではなく,最適化手法を用いて攻撃を行うため,生成される敵対的サンプルの特徴が異なる可能性がある.そのため,提案手法による防御効果が特に高くなったと考えられる.