データ多様体の埋め込み幾何学に基づく新しい敵対攻撃法の提案

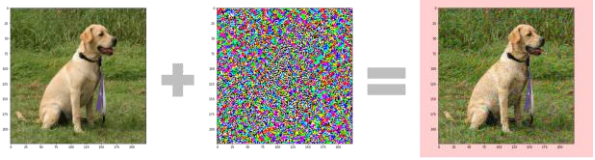
Novel Adversarial Attacks Based on Embedding Geometry of Data Manifolds

森田 匡博・ネットワーク分科会・中央大学

It has been shown recently that adversarial examples inducing misclassification by deep neural networks exist in the orthogonal complementary spaces of the tangent spaces of the data manifold. In this paper, we propose novel adversarial attacks based on the embedding geometry of the data manifold. The proposed attacks generate adversarial examples by adding imperceptible perturbations in the directions of the orthogonal complementary space of the tangent spaces of the data manifold along which the weight vectors have prominent components. Moreover, we also consider targeted attacks by the output inversion in the hidden layer neurons toward the target class. Evaluations of these proposed attacks are also reported.

1.研究背景・概要

深層学習を利用した画像認識などでは，人間が知覚できないほど小さな摂動を加えて生成される敵対的サンプルによって，誤分類を引き起こすことが発見されている．



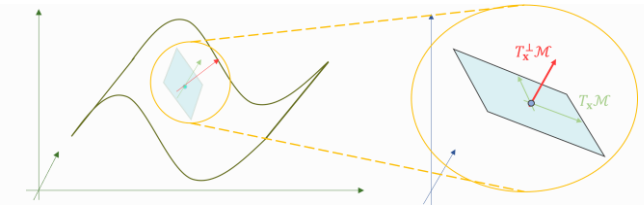
最近学習データが持つデータ多様体の埋め込み構造を解析することで，敵対的サンプルはデータ多様体の接空間の直交補空間方向に存在することが明らかにされた*．

本研究では，上述の発生メカニズムに基づき，埋め込み空間におけるデータ多様体構造に着目した新しい敵対的サンプルの生成手法を提案し，これらの攻撃可能性についての評価を行う．

2.先行研究*

データ集合は埋め込み空間$\mathcal{S}$に存在し，低次元の部分多様体$\mathcal{M}$構造を持つ．$n+1$次元の射影空間を考えると，$\mathcal{M}$上のある点$\mathbf{x}$における接空間と直交補空間への直交分解が成立する．

$$T_{\mathbf{x}}\mathcal{S} = T_{\mathbf{x}}\mathcal{M} \oplus T_{\mathbf{x}}^{\perp}\mathcal{M}$$



よって，重み$\mathbf{w}$，摂動$\mathbf{r}$が載った敵対的サンプル$\tilde{\mathbf{x}}$は各成分に分解でき，NNに入力された際の，重みとの内積は以下の式で表せる．

$$\mathbf{w}^{\top}\tilde{\mathbf{x}} = \mathbf{w}_{\mathcal{M}}^{\top}\mathbf{x}_{\mathcal{M}} + \mathbf{w}_{\mathcal{M}}^{\top}\mathbf{r}_{\mathcal{M}} + \boxed{\left(\mathbf{w}_{\mathcal{M}}^{\perp}\right)^{\top}\mathbf{r}_{\mathcal{M}}^{\perp}}$$

多様体方向の成分$\mathbf{r}_{\mathcal{M}}$は，正常入力間の変形を表すため，人間に気づかれないと定義される敵対的サンプルにはほぼ含まれない．
したがって，誤分類は直交補空間方向の摂動$\mathbf{r}_{\mathcal{M}}^{\perp}$によって引き起こされる．

3.提案手法

学習データが持つデータ多様体の埋め込み構造に基づく，データ多様体の直交補空間方向に対応したニューラルネットワークの重みを活用した敵対的サンプルの生成手法を提案する．（一部抜粋）

➢ 攻撃法I

$\left(\mathbf{w}_{\mathcal{M}}^{\perp}\right)^{\top}\mathbf{r}_{\mathcal{M}}^{\perp}$ を最大化させるために，摂動を重みの直交補空間方向とする手法．次式で計算することができる．（$\mathbf{W}_{\mathcal{M}}^{\perp}$は列ベクトルは多様体の直交補空間方向へ射影された重みの直交補空間ベクトル全体，hは1層目の中間層のニューロン数，$\mathbf{1}_h$は全要素が1のh次元列ベクトル）

$$\mathbf{r} = \mathbf{W}_{\mathcal{M}}^{\perp}\mathbf{1}_h$$

➢ 攻撃法II

事前の解析にて，重みと正常画像の内積をシグモイド関数に与えた結果，0.0あるいは1.0付近に集中．
→中間層の出力が反転するような重みの直交補空間方向を用いて摂動を生成する手法を考える．

$$\mathbf{r} = \begin{cases} \mathbf{r}_+ & \text{(1)0.2以下を取るニューロンを反転} \\ \mathbf{r}_+ - \mathbf{r}_- & \text{(2)0.2以下&0.8以上を取るニューロンを反転} \\ -\mathbf{r}_- & \text{(3) 0.8以上を取るニューロンを反転} \end{cases}$$

例（(2)の場合）:

4.実験と結果

【実験】

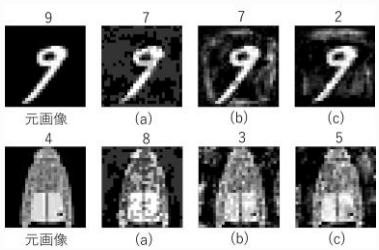
本研究では，提案手法をFGSMとの比較により検証を行う．MNIST内の訓練データ中の1万枚のデータセットおよびFashion-MNIST内の訓練データ中の1万枚のデータセットを対象に攻撃画像を生成する．

【結果（一部抜粋）】

1．MNISTでの攻撃成功率

攻撃手法	攻撃成功率(%)			
$\epsilon =$	0.05	0.1	0.15	0.2
FGSM	53.82	89.81	97.83	99.42
攻撃法I	19.05	47.75	63.33	73.66
攻撃法II-(1)	33.31	83.04	96.90	99.68
攻撃法II-(2)	20.03	68.40	88.47	95.78
攻撃法II-(3)	5.07	34.23	60.90	77.41

2．敵対的サンプルの描画



(a)FGSM
(b)攻撃法I
(c)攻撃法II-(2)

を用いて生成された敵対的サンプル画像

5.結論と今後の課題

【結論】

本研究では，データ多様体の直交補空間方向の摂動により誤分類が引き起こされるという発生メカニズムに基づき，既存手法とは全く異なる新しい敵対的サンプルの生成手法を提案した．データの変形に影響が少ない直交補空間方向のみを摂動とすることにより，元画像の見た目への影響を抑えた敵対的サンプルが生成されることを確認した．

【今後の課題】

- 生成される摂動の可視性の評価方法の確立
- その基準に従った攻撃方法の検討

*H. Tasaki, Y. Kaneko, and J. Chao, "Curse of co-dimensionality: Explaining adversarial examples by embedding geometry of data manifold"