データ多様体の埋め込み幾何学と当てはめに基づいた深層学習における敵対的サンプル攻撃の検知手法

Detecting Adversarial Example Attack Based on Embedding Geometry and Fitting of Data Manifold

久重広樹・ネットワーク分科会・中央大学

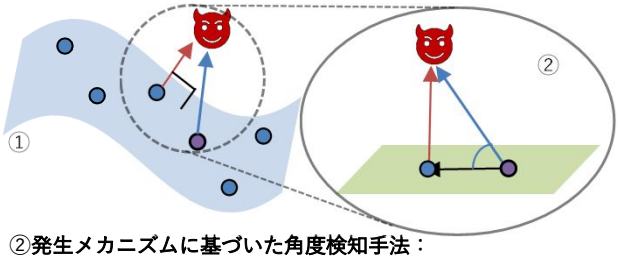
Abstract - However neural networks have contributed to many fields, Adversarial Example Attack that causes misclassification has been discovered. This attack data is generated by adding perturbations to normal data and both are so similar to each other that we cannot distinguish. The vulnerability of neural networks has been widely recognized as a problem and some countermeasures have been limited in their effectiveness because they do not consider the attack mechanism. In a previous study, they proposed a detection method based on the orthogonal complementary space component of the data manifold as a mechanism for generating hostile samples, but the detection accuracy decreased when the density of the data set was small. In this paper, we propose a quadratic fitting method of the data manifold and improve the detection accuracy of the related method based on the mechanism.

1. 研究背景

・AIに対する攻撃（敵対的サンプル）が発見された.
・既存の対策手法は効果が限定的である.
・それは，攻撃の発生原因に基づいていないためである.

2. 先行研究

①攻撃発生メカニズムの解明：
・攻撃データはデータ多様体（分布）の外に存在する.
　→データ多様体の直交補空間成分を持つ.



②発生メカニズムに基づいた角度検知手法：
・データ多様体と入力点との角度を用いて検知する.

【課題】
・入力攻撃の元データが未知であるとき検知精度が低下する.
　データセットの密度が低いことで最近点が遠く，
　角度が上手く計算できないため.

3. 提案手法（二次当てはめ）

・低密度離散データセットに対して連続な2次当てはめを計算し，幾何学的な計算を可能にする.
・発生メカニズムに基づいた既存対策手法の検知精度向上を図る.

①データ多様体の近傍ごとに二次**当てはめ**を計算.
・入力点$\boldsymbol{x}$の最近点$\boldsymbol{x}_0$を中心にk-近傍$N(\boldsymbol{x}_0)$を作成する.
・k-近傍$N(\boldsymbol{x}_0)$にPCAを適用し，特徴空間基底Uを得る.
・k-近傍を$U^{-1}(N(\boldsymbol{x}_0)-\boldsymbol{x}_0)$で座標変換する.
・以下，誤差が一定値以下になるまでdを増やしていく.
　- Uの先頭からd個目までを**入力基底**とし，データ点$\mathbf{x}_i$を分割.

$$\mathbf{x}_i=(\mathbf{x},\mathbf{x}')^{\mathrm{T}}=\left(\mathrm{x}_1,\dots,\mathrm{x}_d,\mathrm{x}_{d+1},\dots,\mathrm{x}_{\min(n,k-1)}\right)^{\mathrm{T}}\in N(\boldsymbol{x})$$

　- $j(=d+1,\dots,n)$次元目の当てはめy_jを計算する関数α_jは

$$\mathrm{y}_j=\alpha_j(\mathbf{x})=\mathbf{x}^{\mathrm{T}}A\mathbf{x}+\boldsymbol{b}\mathbf{x}+c.$$

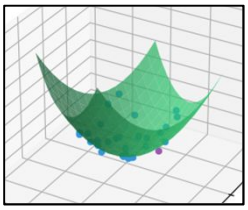
　- 当てはめ関数α_jのパラメタは最小二乗法で一意に定まる.

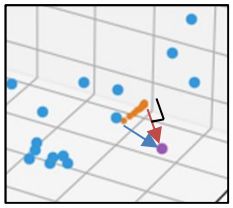
②当てはめ関数に勾配法を用いて最近点を再探索.
・最近点を座標変換$\mathbf{x}_0=U^{-1}(\boldsymbol{x}_0-\boldsymbol{x}_0)$し勾配法の初期値とする.

$$\mathbf{z}_0=\mathbf{y}_0=\boldsymbol{\alpha}(\mathbf{x}_0)=\boldsymbol{c}$$

$$\mathbf{z}_{\mathrm{iter}}=\mathbf{z}_{\mathrm{iter}-1}-\delta\frac{\partial E}{\partial \mathbf{z}}$$

・ここでEを入力点$\boldsymbol{x}$と$\mathbf{z}_{\mathrm{iter}}$の二乗和距離関数とした.


①の可視化


②の可視化

①：ガウスノイズを加えた離散データセットから，
　連続な関数を推定できた．青：データセット，紫：入力点
②：元の最近点から当てはめ関数上を推移して，
　より近い点を計算できた．橙色：勾配法の推移

4. 実験

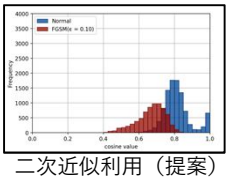
①画像データセットにおける最近点の再計算
・攻撃入力に対してデータセット上の最近点と，
　提案手法における二次関数上の再計算最近点を比較した.
【結果】
・再計算最近点は入力との距離を短くした．（5.9→3.5）
・入力画像の特徴（ループの最後の繋がり）を保存し，
　攻撃画像の特徴（背景の雑音）も除去した.

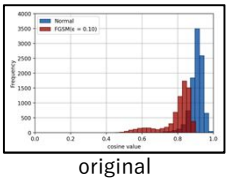

入力（攻撃）

最近点

最近点（提案）

②既存検知手法への応用
・最近点のズレを解消することを目的に二次近似を用いた手法と既存手法におけるコサイン類似度を比較した.
【結果】
・検知精度が6%上昇した．（60%→66%）
・両方の角度が直交したため，新しい検知指標が必要である.


二次近似利用（提案）


original