深層学習における敵対的サンプル攻撃に対するデータ多様体補間を用いた検知手法に関する研究

Detecting Adversarial Example Attacks in Deep Learning by Data Manifold Interpolation

藤田真緒・暗号・認証分科会・中央大学大学院

Abstract - Deep learning is one of the machine learnings used in image recognition field. However, there has been an attack called “Adversarial Example Attack” lately, which causes a misclassification by neural networks by inserting an imperceptible perturbation. There are multiple hypotheses of why adversarial examples exist, and new hypothesis which is based on data manifold structure has been proposed recently. In this study, we apply a data manifold interpolation and data augmentation to existing angle detection.

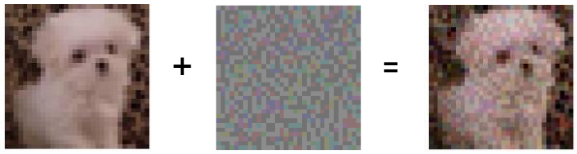
研究背景 / 目的

敵対的サンプル(AE)攻撃：画像に微小な摂動を加えることにより機械学習モデルの誤分類を引き起こす攻撃

・AEはデータ多様体構造における接空間の直交補空間成分を持つ．

・検知手法のひとつにこの性質を利用した角度検知手法がある．

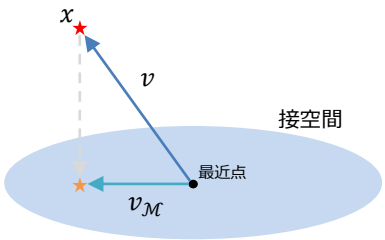
→データ多様体補間とデータ拡張のMixupを用いた角度検知による検知精度の向上



先行研究：角度検知手法

・AEの性質を利用し，接空間に対する正常データとAEの存在位置の違いから検知する手法

1.入力点xの最近点をデータ多様体$\mathcal{M}$から選択し，最近点のk近傍を作成する．
2.k近傍にPCAを適用し，接空間基底を求める．
3.最近点から入力点xへのベクトルv，接空間への射影ベクトルを$v_{\mathcal{M}}$とし，vと$v_{\mathcal{M}}$のコサイン類似度を求める．
コサイン類似度$\frac{v \cdot v_{\mathcal{M}}}{\|v\| \cdot \|v_{\mathcal{M}}\|}$が小さいとき，敵対的サンプルであると考えられる．



提案手法I：二次多様体補間

・当てはめを用いて最近点と接空間の再計算を行うことで，より正確な接空間を推定する手法

1.入力点の最近点の近傍に二次当てはめを行い，求めた当てはめ関数からデータの推定を行う．
2.入力点とある点の距離関数に最急降下法を適用し，最近点を再計算する(初期値：最近点)．

提案手法II：近傍内のMixup

Mixup：二枚の画像を合成して新たな画像を作成するデータ拡張の一種

・近傍内のMixupを行いデータ密度を向上させることで，より正確な接空間を推定する手法

1.入力点の最近点のk'近傍を作成する．
(最終的に作成する近傍よりも広く近傍をとる)
2. k'近傍の画像を同じラベル同士でMixupし，近傍を拡張する．
3.拡張した近傍から再度，最近点のk近傍を作成する．

実験結果

実験設定： データセット：CIFAR10
モデル：CNN(分類精度81.51%)
データ多様体：入力データ

結果と考察

近傍数	従来法	手法 I	手法 I+II
k=100	49.25%	51.40%	53.87%
k=200	69.48	68.59%	70.84%

表：近傍数の違いによる検知精度

提案手法を適用することにより，接空間の推定精度が上がり，検知精度が向上した．また，近傍数が多い程より正確な接空間が作成できると考えられる．

今後の課題

入力データでデータ多様体を構成した際の近傍は近傍内のラベルがばらける傾向にある．一方で，CNNの中間層の出力を構成に用いると近傍内のラベルはひとつに集中する傾向があるため，この近傍のラベル分布を生かした新たな検知手法の提案を行いたい．