深層学習における敵対的サンプルとデータ拡張による データ多様体の接空間推定

Defense Adversarial Examples Based on Embedding Geometry of Data Manifolds and Estimate the Tangent Space of the Data Manifolds using Data Augmentation

宮坂優吾・ネットワーク分科会・中央大学

1.研究背景

近年、機械学習の応用範囲は広がり、様々な分野で成果を上げている。
機械学習には敵対的サンプルと呼ばれる特殊な入力に対して脆弱性があることが知られている。
敵対的サンプルは正常データで作られた多様体(データ多様体)とは別の多様体に存在する。
敵対的サンプルの発生メカニズムに基づく防御手法を提案された。

2.提案手法

敵対的サンプルの発生メカニズムに基づく防御手法の中では、正常データによるデータ多様体の接空間の推定が行われている。
現状、データ密度が低いデータセットでは接空間の推定がうまくいっていない。
そこで接空間を推定する際、正常データのデータ拡張を行い、より正確な接空間の推定に試みた。

3.結果、今後の課題

正常データを2種類の方法で拡張することで防御後の分類精度を最大約2%向上した。
今後は、より正確な推定を行うことができるデータ拡張方法を探す。

拡張倍率	1倍	3倍 縦横シフト	3倍 回転
K=50	32.80%	34.66%	33.55%
K=150	35.72%	37.12%	36.32%