データ多様体構造に着目した敵対的攻撃の防御手法に関する研究

Defending Adversarial Examples Based on Embedding Geometry of Data Manifolds

藤田真緒・暗号・認証分科会・中央大学大学院

研究背景

深層学習は、自動運転や生体認証などの画像認識技術に広く用いられている。しかし、画像認識分野には、敵対的サンプルと呼ばれるAIの脆弱性が存在する。敵対的サンプルの存在原因は、様々な仮説が立てられているが、未だ敵対的サンプルの性質の完全な解明は行われていない。

研究目的

2022年に新たな仮説として、データの多様体構造に着目したものが発表された。現在、この仮説に基づいた防御手法が提案されているが、さらなる精度の向上した防御手法の提案を目的とする。

今後の方針

ニューラルネットワークの重みに着目した防御手法の提案を行う。