AIを悪意のある攻撃から守れ
～AIセキュリティ、敵対的攻撃の検出手法に関する研究～

広島市立大学大学院情報科学研究科知能工学専攻
データ科学講座・データ工学グループ・AIセキュリティチーム

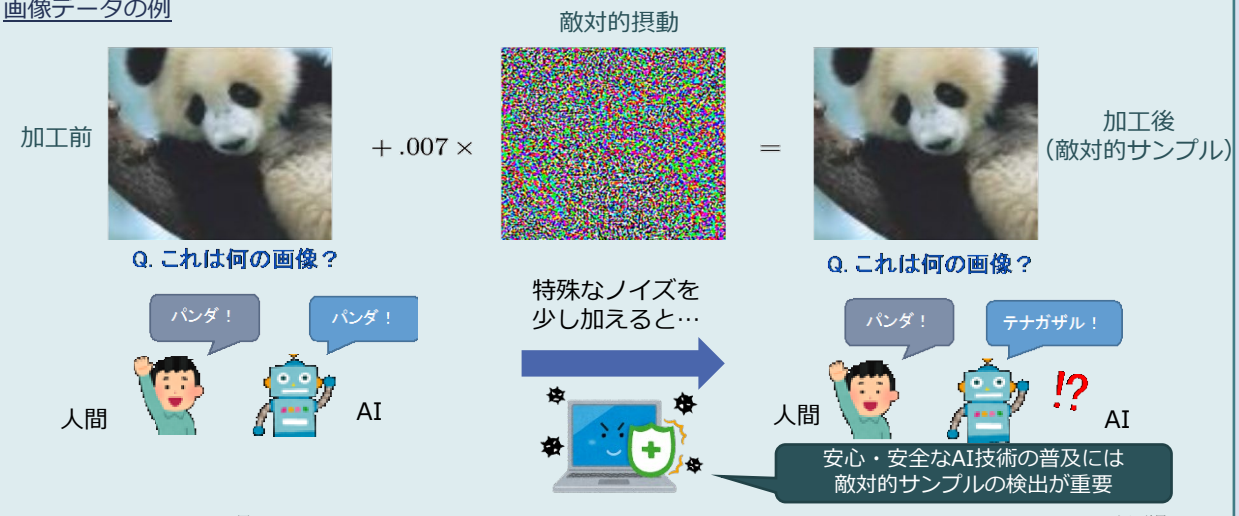
AIセキュリティへの関心が高まる中、データに悪意のある細工を施して深層学習のモデルを騙す敵対的攻撃が問題になっています。本展示では、音声データや時系列データに対する敵対的攻撃とそれらを検出する手法について紹介します。

敵対的サンプル

敵対的サンプル （AE: Adversarial Example）

➡ 人間には認識できないような微小なノイズでAI（深層学習モデル）の誤認識を引き起こすデータ

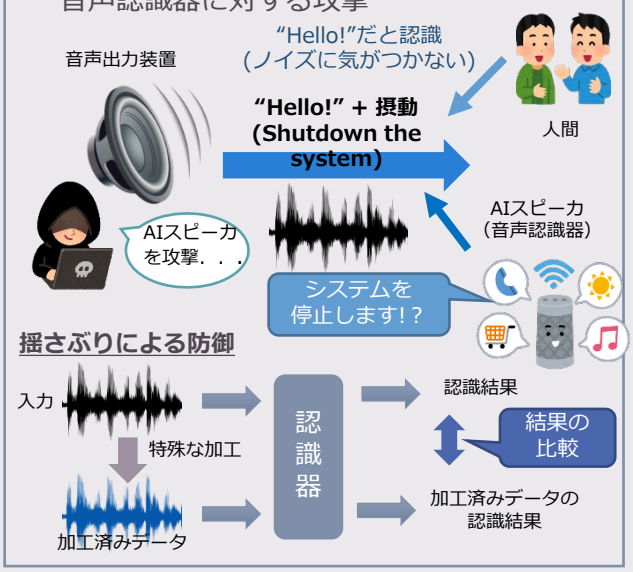
画像データの例



図はI. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in International Conference on Learning Representations, 2015. より引用

音声データ

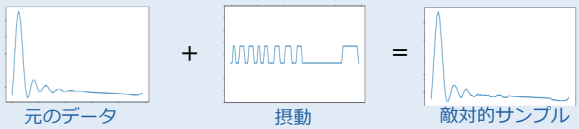
AAE: Audio Adversarial Example
音声認識器に対する攻撃



時系列データ

時系列データ分類器に対する敵対的サンプル

心電図検査，地震検知，食品検査などに使用



検出モデルによる防御

敵対的サンプルかどうかを判定するAI（=検出モデル）を使用