見えない攻撃から、AIを守れ
～AIセキュリティに関する研究～

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・AIセキュリティグループ

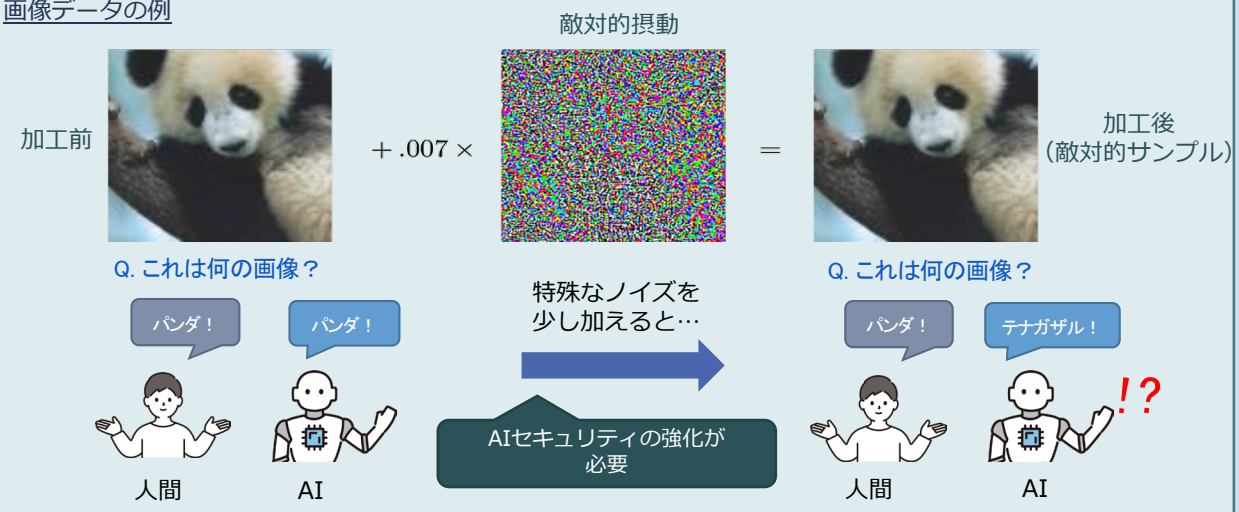
人間には気づきにくい小さなノイズや加工によって、AIが誤認識してしまう攻撃が問題となっています。本研究では、音声認識AIや自動運転向けAIを対象に、入力データを少し変化させたときの認識結果の不安定さを手がかりとして、攻撃を検出する技術を研究しています。

敵対的攻撃

敵対的サンプル （AE: Adversarial Example）を用いて深層学習モデルを騙す攻撃

➡ 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

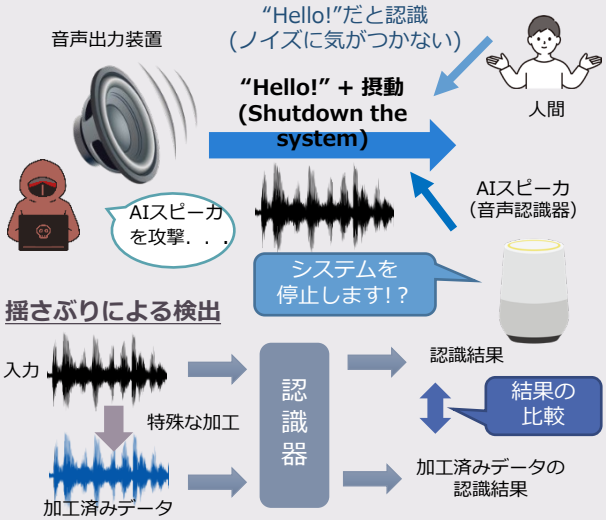
画像データの例



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



入力を加工し、認識結果が大きく変わるか調べる

３次元点群データ

３次元点群データに対する敵対的サンプル

LiDARや深度カメラを用いた物体認識に対する攻撃

差分計算を用いた生成