決定境界に基づくブラックボックス音声認識モデルへの敵対的サンプル攻撃
Adversarial Black-Box Attacks Based on Decision Boundary for Audio Recognition

山本恭敬・暗号・認証分科会・中央大学大学院

研究背景

- 近年, 音声認識AIが広く普及している.
- 画像認識, 音声認識, 自然言語処理などで敵対的サンプル攻撃というAIを騙す攻撃が確認されている.
- サーベイ論文 [Cheng+22]によると, 音声認識モデルへのブラックボックス攻撃において決定境界に基づくものはなかった.

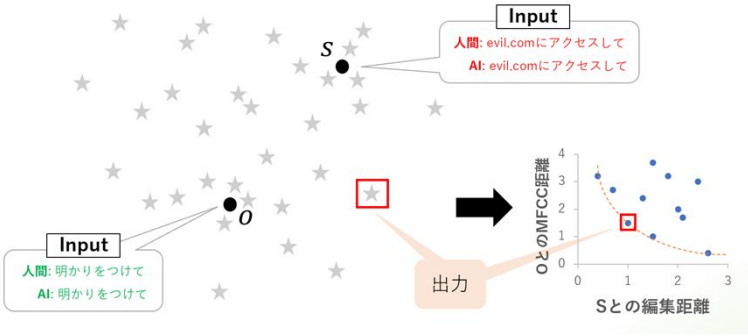
現在の進捗状況

- Boundary Attack [Brendel+18]は画像認識モデルへのブラックボックス攻撃である。この手法から着想を得て決定境界に基づく攻撃手法を考案・実装し, 攻撃の有効性を確認した.

今後の予定

- 聞き取り調査などの評価を行う.
- より実世界の制約を考慮した攻撃を考案する.

既存手法：多目的最適化 [Khare+19]



提案手法：決定境界に基づく攻撃