ブラックボックス音声認識モデルに対する敵対的サンプル生成の新手法

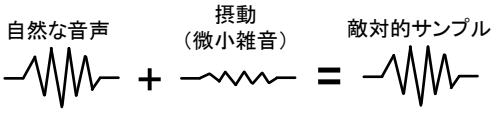
New Method of Generating Adversarial Examples for Black-Box Audio Recognition Model

山本恭敬・ネットワーク分科会・中央大学大学院

Abstruct - One of the methods of attacking machine learning automatic recognition systems is the adversarial sample attack. This refers to a method that generates samples such that the human recognition and the machine learning model's decision differ, and uses these samples to attack the system. In previous research, there has been little research on the generation methods of adversarial sample generation for speech recognition models, which are classified as black box attacks and targeted attacks. Therefore, we propose a new method of generating samples close to the decision boundary as a new method of adversarial sample generation for speech recognition models. This method is a black box attack that can generate samples without knowing the details of the speech recognition model, and is classified as a targeted attack that generates samples such that the speech recognition model recognizes the speech as the attacker intended. In order to verify the effectiveness of the proposed method, we generate adversarial samples against speech recognition models such as Whisper and confirm that the proposed method can generate higher quality samples than those generated by existing methods.

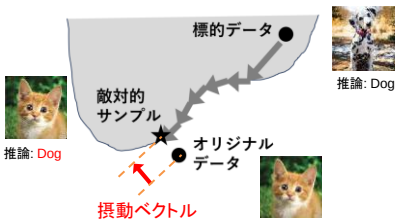
背景

- 人間の自然な知覚とAIの推論が異なるようなデータを**敵対的サンプル**といい，AIの脆弱性として知られている．



人間	こんにちは	こんにちは
AI	こんにちは	さようなら

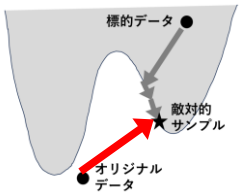
- **境界攻撃**は画像の敵対的サンプル生成手法として2018年にBrendelらによって提案された．
- オリジナルデータと標的データを入力し，敵対的サンプルを出力する．
- 標的データのラベル（標的ラベル）と同じクラスの範囲で，オリジナルデータに貪欲に近づくように微小移動を繰り返して敵対的サンプルを生成する．
- 攻撃対象モデルのパラメータが不明でも実行できる．（ブラックボックス攻撃）



※犬と猫の画像は論文[Brendel+2018]より引用

課題

- 境界攻撃による音声の敵対的サンプル生成を試みたが，生成音声を聞いたときに標的データの音声が混ざって聞こえてしまった．
- 画像認識モデルよりも音声認識モデルのほうが決定境界が複雑であること，および一般に決定境界が複雑だとオリジナルデータから遠い局所解に収束してしまうことが原因と推測できる．

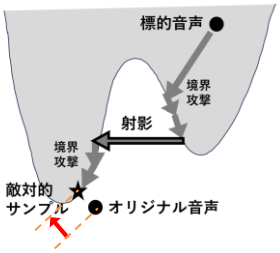


提案手法

オリジナルデータに近づきつつ局所解から脱出する手法として以下の**射影アルゴリズム**を考案．

入力：射影前の音声 a，オリジナル音声 o，標的ラベル T．
出力：射影後の音声．
① $i \leftarrow 1, n \leftarrow (a$ の要素数$)$．
② 音声 $p = (a_1, a_2, \ldots, a_{i-1}, o_i, a_{i+1}, \ldots, a_{n-1}, a_n)$ とし，音声認識モデルによる p の文字起こしが T に一致したら $a \leftarrow p$．
③ $i < n$ ならばを i を1進めて②へ，そうでなければ a を出力して終了．

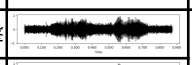
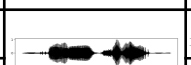
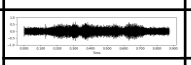
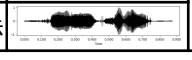
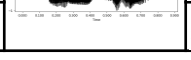
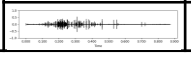
境界攻撃と射影を交互に繰り返すことで，境界攻撃のみの場合と比べてよりオリジナルデータに近い敵対的サンプルの生成が期待できる．



音声試聴はこちらから

実験結果

表　境界攻撃により生成した敵対的サンプルと提案手法により生成した敵対的サンプル．（モデル呼び出し回数は同一にそろえた．）

	敵対的サンプル a	オリジナルデータ o	摂動 $a-o$	$\lvert a-o \rvert$
境界攻撃				18.6
提案手法				**2.46**

- 上表より，提案手法のほうがオリジナルデータにより近い敵対的サンプルを生成できている．
- 右図より，文字誤り率の平均値（×印）や中央値（点線）が低い．これは実験参加者による文字起こしが攻撃者の意図通りのテキストに近いことを表す．
- 音声認識モデルにwhisper [Radford+2023]を使用した．

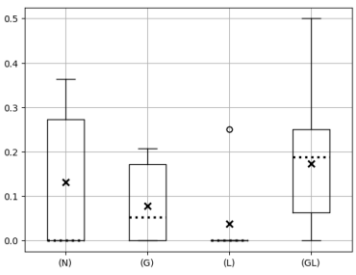


図　提案手法で生成した4つの敵対的サンプルの実験参加者による文字起こしテキストと書く敵対的サンプルに対応するオリジナルデータのラベルテキストの文字誤り率．

今後の課題

検知アルゴリズムをかいくぐる敵対的サンプル生成手法の開発．
（モデル呼び出し回数の減少，スパイクノイズの減少など）