人間の視覚特性に合わせたAdversarial Perturbationの最小化手法

Minimization of Adversarial Perturbation Adapted to Human Visual Characteristics

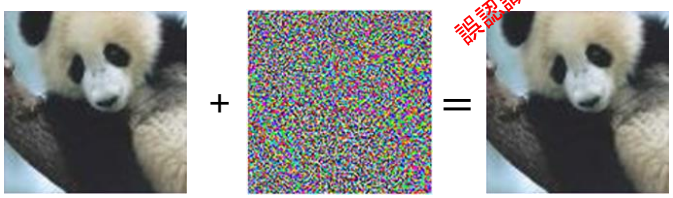
近藤大生・システム分科会・情報セキュリティ大学院大学

Abstract

Though Deep Neural Networks(DNNs) are well acknowledged as the best image scene recognizer and are widely used in real world systems and applications, DNNs are also known to have a vulnerability to Adversarial Example(AE) attack : image forgery by adding fine perturbations to legitimate data. AE may cause malfunction in DNNs. This can potentially lead to serious consequence especially in security-sensitive applications. Although various attack methods have been proposed at present, there is no AE that evaluates the conspicuousness of noise. Therefore, some of the proposed attack methods have samples with prominent noise . In this paper, we propose a new attack method adapt to three Human Visual Characteristics : ″Ability of color identification″, ″Area of focus″ and ″Ability of frequency discrimination″. As a result of experiments, we found that the AE by ″Ability of frequency discrimination″ is the most natural noise for humans.

1. 研究背景| 敵対的サンプルの脅威

ディープニューラルネットワーク(DNN)特有の誤認識問題

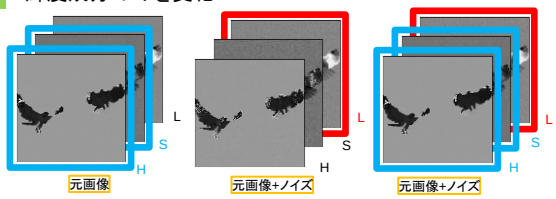


Lpノルムによるノイズ制限は
必ずしも人間が見て違和感のないノイズになるとは限らない

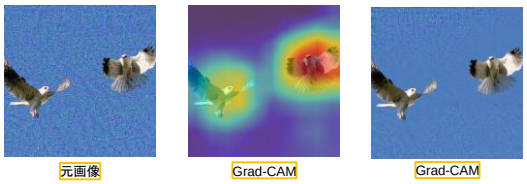
目的:人間の視覚特性を加味した敵対的サンプルの作成

2. 提案手法|人間の視覚特性に準じたAE作成

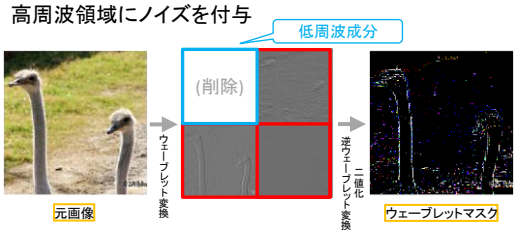
色　輝度成分のみを変化



場所　背景に大きなノイズを付与



周波数　高周波領域にノイズを付与

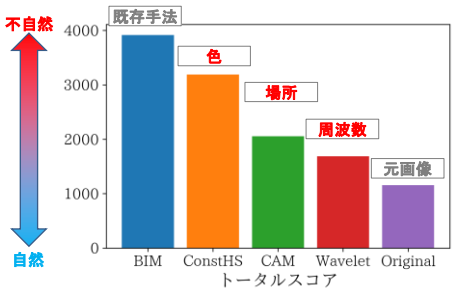


3. 生成した敵対的サンプル

BIM(既存手法)

constHS　CAM　Wavelet

4. 提案手法による敵対的サンプルの視覚評価



既存手法と比べ目立たないノイズを生成することができた.

5. まとめ

- 従来手法と比べ視覚検出が困難な敵対的サンプルを作成した.
- 人間がどのようなノイズに弱いのかを示した.
  - 高周波領域
  - 背景
  - 明るさのみの変化