AIモデル説明Grad-CAMに対する敵対的攻撃の提案と評価

明治大学
寶木　隆正　　菊池　浩明

研究背景

- 深層学習モデルのブラックボックス性と説明可能AI（XAI）の重要性
  ハスキー vs 狼問題
  モデルは本質的特徴ではなく背景に依存していた
  モデルの判断根拠を可視化・解釈するXAIが重要となる。

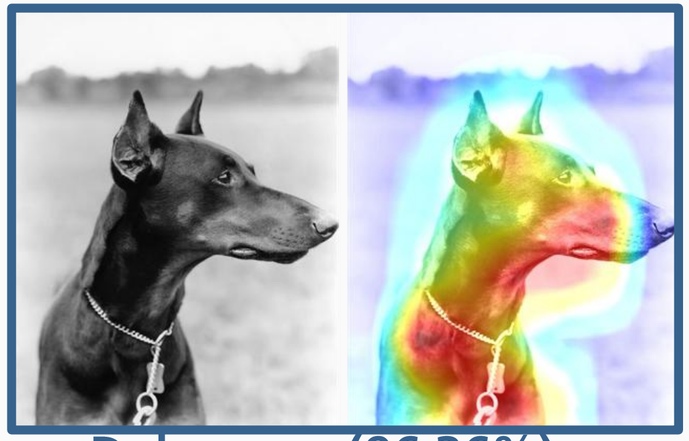
- 特にGrad-CAMなどが注目されている

(a) Husky classified as wolf　(b) Explanation

出典 https://arxiv.org/abs/1602.04938

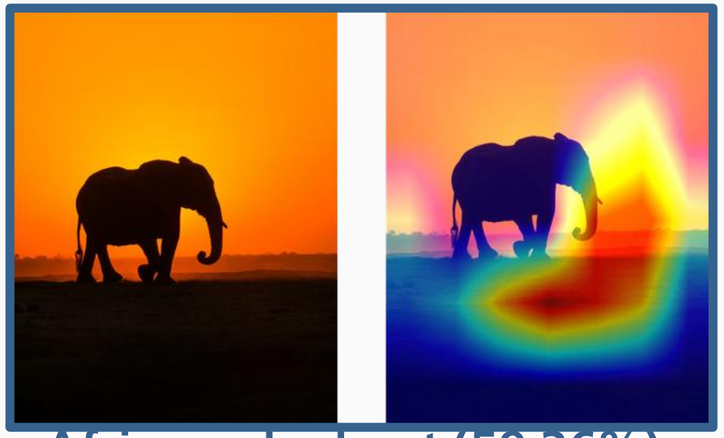
予備実験（100枚のイメージのXAIの信頼性）

A 成功例



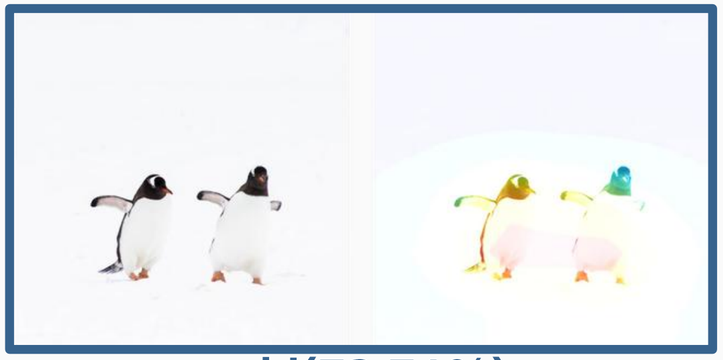
Doberman (96.36%)

B 分類は正しいがXAIが誤ってる



African_elephant (59.26%)

C XAIは正しいが誤分類



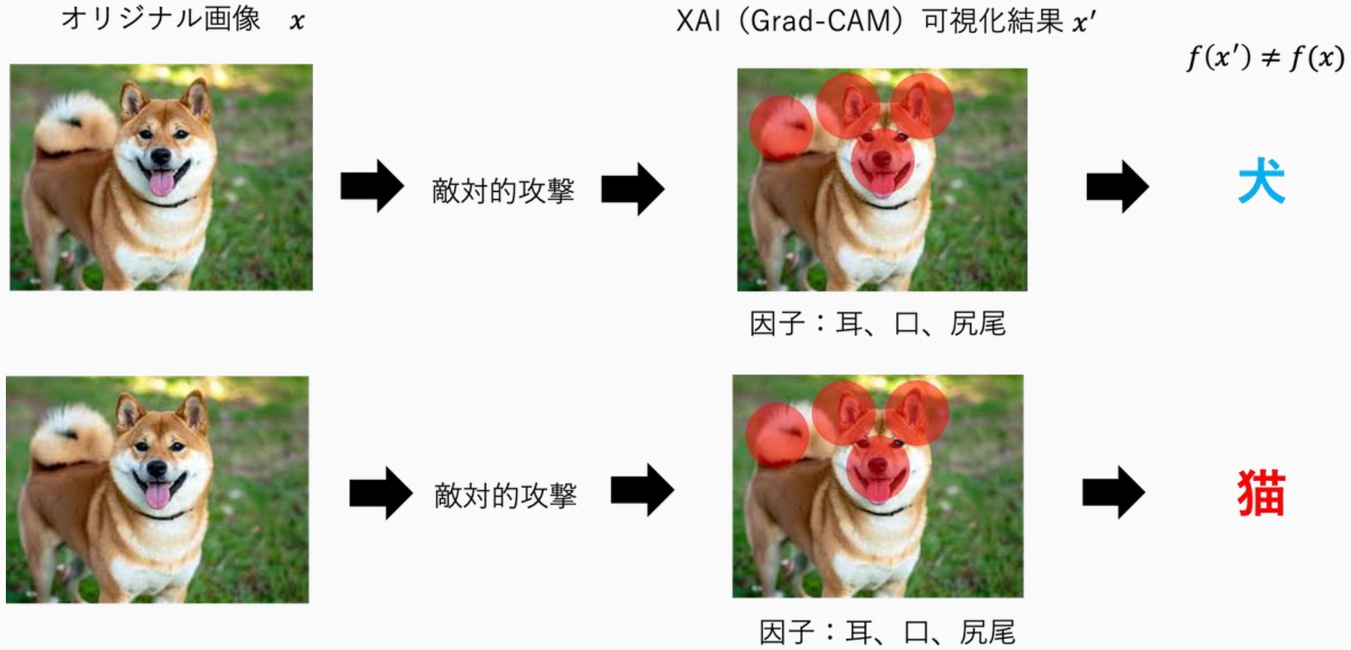
ski(72.71%)

研究動機

Cのパターンを人工的に
作れるのか？

研究目的

Grad-CAMの可視化結果を維持したまま、モデルを誤分類させる攻撃の検証



先行研

究 XAI

Grad-CAM: CNNの最後の畳み込み層の勾配情報を用い、クラス判断に影響した画像部位を可視化する手法

Selvaraju, R. R., et al. “Grad-CAM: Visual Explana-tions from Deep Networks via Gradient-Based Local-ization.” ICCV, pp.4-5,2017.

敵対的攻撃

FGSM (Fast Gradient Sign Method): ニューラルネットワークの勾配を利用して入力画像に摂動（ノイズ）を加え、誤分類を引き起こす標準的な攻撃手法

Goodfellow, I. J., et al. “Explaining and HarnessingAdversarial Examples,” ICLR, pp.2-3,2015.

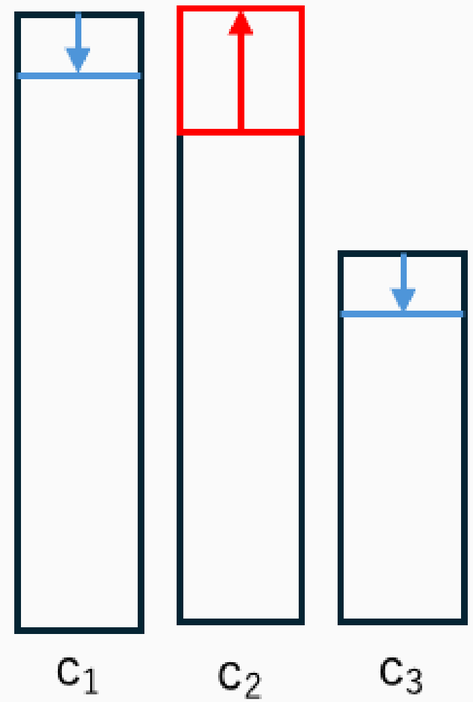
解決方法

予測値の「順位」C_1,C_2,C_3に注目したホワイトボックス攻撃

従来の交差エントロピーではなく、モデルが出力する予測値(Z)の順位を操作する

提案1　FGSML1：$L_1(\theta, x, y) = z_{c2} - z_{c1}$

提案2　FGSML2：$L_2(\theta, x, y) = (z_{c2} - z_{c1}) + (z_{c2} - z_{c3})$



実験方

法 データセット

ImageNetからサンプリングした1000枚の画像のうち、
攻撃前に正しく分類された865枚を評価対象として使用した

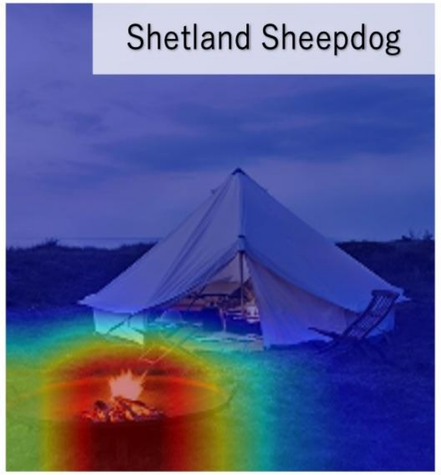
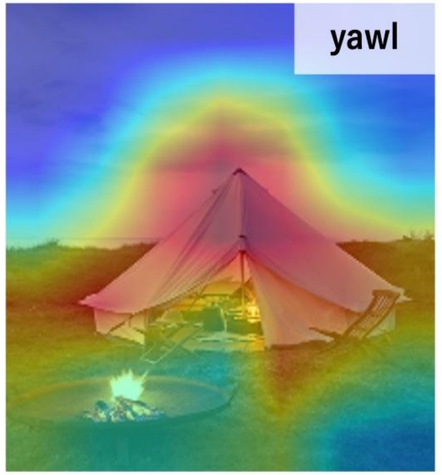
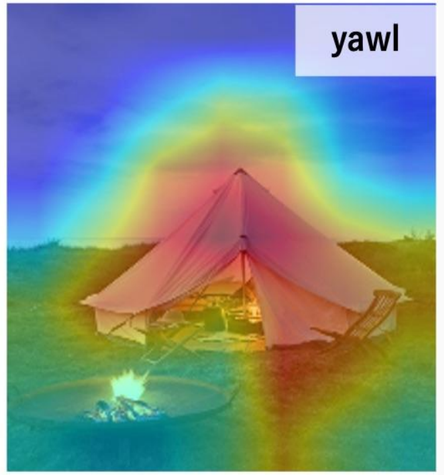
対象モデル

ResNet50

評価

- **Attack Success Rate (ASR)**
  攻撃によって1位の予測クラスが変化した確率
- **Mean Observed Dissimilarity (MOD)** $MOD = \frac{1}{N}\sum_{i=1}^{N}(1 - SSIM(img_i^{original}, img_i^{Adversarial}))$
  XAIのモデル説明の変化量を示す指標。0に近いほどモデルの注目領域がほぼ同一で、1に近いほど注目領域が異なることを示す

結果①

分類 XAI	失敗 変化	失敗 不変	失敗 不変
			
Original	FGSM_CE	FGSM_L1	FGSM_L2
1-SSIM	0.415	0.176	0.118

結果②

攻撃成功率とXAI モデル説明変化

攻撃手法	攻撃成功率 ASR（%）	平均類似度(1-SSIM) MOD
FGSM_CE（従来）	78.03	0.110
FGSM_L1（提案）	80.00	0.099
FGSM_L2（提案）	74.34	0.086

分類成否と可視化変化の割合（%）

分類	成功	成功	失敗	失敗
XAI	不変	変化	不変	変化
FGSM_CE（従来）	19.54	2.43	34.80	43.24
FGSM_L1（提案）	18.38	1.62	42.66	37.34
FGSM_L2（提案）	24.62	1.04	44.28	30.06

考察

なぜ注視点を動かさずに誤分類が可能なのか

提案手法は特定の予測順位（特に2位）への誤分類を狙い撃ちする

AIの判断における注目領域を大きく変化することなく、確信度を操作して誤分類を誘発できたと考えられる

結論

- **XAIの注目点を動かさずに高い確信度で誤分類を引き起こすことを実証した。**

- **今後の課題**

  PGDなどのより強力な反復攻撃への適用。
  ResNet以外のモデルに対する検証。
  モデル分類を変えずに注目点を動かすことができるか。

分類
XAI

Original

失敗
不変

FGSM_CE

失敗
不変

FGSM_L1

失敗
不変

FGSM_L2

分類
XAI

Original

成功
変化

FGSM_CE

成功
変化

FGSM_L1

成功
変化

FGSM_L2