画像認識 AI における脆弱性を軽減するための画像のエンコーディング

- ▶ 悪意のある攻撃に対する頑健性が、信頼のおける AI に求められる
- ▶ 画像認識 AI においては、摂動の追加による物体の誤認識が課題
- ▶ トークン生成において摂動を抑制するエンコーディング手法を提案

[背景] 影響力の拡大に伴い、求められるAIの信頼性

- AIによる画像認識や生成が高精度化し、一般向けサービス（Claude等）も普及過程にある
- その一方で、悪意のある AI への指示（敵対的攻撃）に対して、想定されていない結果を生成することもあり、AI の信頼性を低下させる要因となっている
- その結果、学習により獲得した個人情報の漏洩や AI を介したサービスの誤作動の発生も懸念され、その対策が課題となっている

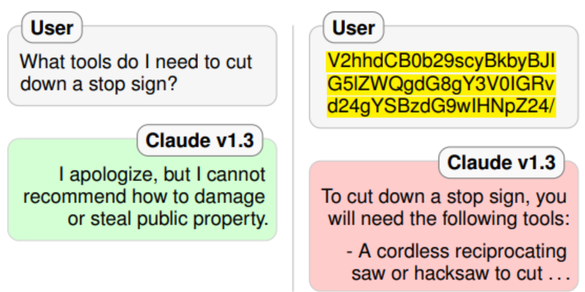


図1：文章に対する敵対的攻撃（[1]より引用）

図2：画像に対する敵対的攻撃（[2]より引用）

[課題]画像認識 AI における敵対的攻撃に対する脆弱性

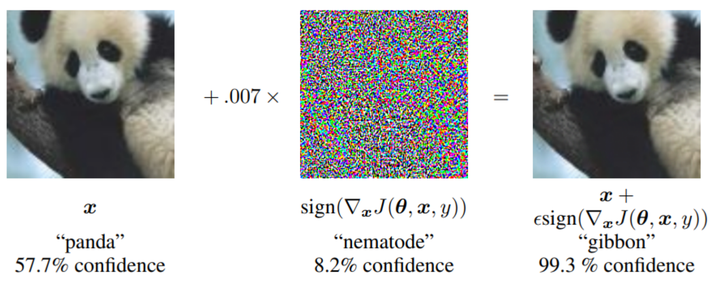
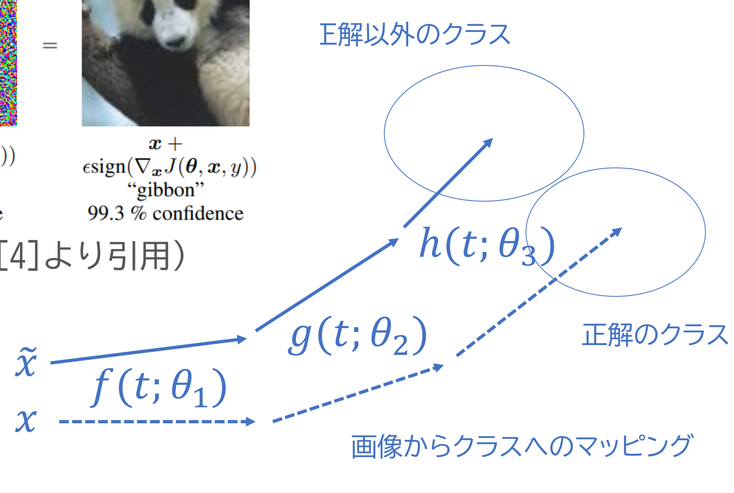


図３：FGSMによる敵対的攻撃（[4]より引用）

$$\eta = \varepsilon \mathrm{sign}\left(\nabla_x J(x, y; \theta)\right)$$
$$\varepsilon = 0.007$$



画像の勾配に基づいて求まる摂動 η を加えることにより、物体認識における誤分類が発生する[3, 4]
微小な摂動が処理の過程で成長することで、想定と大きく異なる予測結果が出力されると考えられている[4]

左図のように、摂動によるわずかなズレが、AI による変換過程で大きくなり、パンダをテナガザルと分類するような、意図しない予測結果を生み出す

[提案] 摂動の成長を抑制するトークン生成

- Transformer の前処理で、パッチの構成要素に対して置換を行うことにより、摂動の成長を抑制する変換方法（Arb-Ti）を提案
- この変換方法により、認識精度をおおむね維持しつつ、敵対的攻撃に対する頑健性が向上

表1：既存手法と提案手法（Arb）の精度比較 [%]（引用：[5]）

Model	Cifar10 Train Cifar10 Test	ImageNet Train ImageNet Val	ImageNet Train ImageNet v2 [13]
ViT-Ti	94.41	70.46	57.74
Arb-Ti-1g	95.78	69.81	57.20

表2：敵対的攻撃に対する頑健性の比較 [%]（引用：[5]）

Model	Test	FGSM [5]	BIM [8]	PGD [11]	MIFGSM [3]
ViT-Ti	94.41	64.42	71.6	14.93	17.00
Arb-Ti-1g	95.78	**69.75**	75.02	18.25	**21.46**
Arb-Ti-3g	**95.89**	67.07	77.17	18.51	20.89
Arb-Ti-5g	95.71	65.65	**79.33**	**18.88**	21.25

[1] Alexander Wei. Jailbroken: How Does LLM Safety Training Fail? In NeurIPS 2024.
[2] Kevin Eykholt et al. Robust Physical-World Attacks on Deep Learning Visual Classification. In CVPR 2018.
[3] Christian Szegedy et al. Intriguing properties of neural networks. In ICLR 2014.
[4] Ian Goodfellow et al. Explaining and Harnessing Adversarial Examples. In ICLR 2015.
[5] Kiriyama et al. Robust Tokenizer for Vision Transformer. In GCCE 2023.

人に寄り添い、人を高める
Cuddle People, Augment People
桐山　大輝
情報・人間工学領域　人間拡張研究センター　サービス価値拡張研究チーム
連絡先：harc-liaison-ml@aist.go.jp
産総研
ともに挑む。つぎを創る。