4U-03

人間がフリーハンドで再現可能な敵対的攻撃

奈良 亮耶† 松井 勇佑‡
東京大学† 東京大学‡

1 はじめに

敵対的攻撃に関するこれまでの研究 [1] は, 人間による解釈のしやすさよりも, どれだけ誤認識を起こせるか, どれくらい知覚しづらいかに主眼を置いてきた. その結果, 従来の攻撃手法によって画像分類器を欺くことに成功しても, モデルが誤分類した理由についての洞察を得ることは難しい. いくつかの研究 [2] は, 敵対的攻撃の解釈可能性について探究している. しかし, 敵対的攻撃と画像分類器の出力の関係に注目して, 解釈可能な攻撃を生成している研究は未だにほとんどない.

私達は, 解釈可能な敵対的攻撃を得るために, 人間がフリーハンドで再現可能な攻撃を提案する. 私達は提案する攻撃手法を, “敵対的落書き” と命名する. 私達は, 入力画像に重ね合わせたときに分類器を欺くために, bézier 曲線の集合を最適化することによって敵対的落書きを生成する. 私達はさらに, 最適化の際に (1) 攻撃の頑健性を高めるために画像のランダム視点変換を導入し (2) 攻撃をよりコンパクトにするために落書き領域の正則化を導入する. 最適化を行った後, 人間が最適化された bézier 曲線の集合を真似して, 入力画像に対して落書きを加える. これにより, 画像分類器に入力画像を誤分類させる.

2 提案手法

本研究では, ベジェ曲線の本数は 6 本で, 一本あたりの制御点の数は 4 個に固定し, ベジェ曲線の色は黒とする. その上で, 入力画像に重ねた際に, 画像分類器がターゲットクラスに誤分類してしまうようなベジェ曲線の集合を求める.

最初にベジェ曲線の制御点の座標を初期化する. 制御点の座標の情報から, Differentiable Rasterizer [3] を用いてベジェ曲線の集合のピクセル情報に変換する. そのピクセル情報を入力画像に重ねたものをモデルに入力する. そしてモデルが入力画像をターゲットクラスに誤分類するようにクロスエントロピー損失をとり, それを元にベジェ曲線の制御点を最適化する. この操作を繰り返す. 私達はさらに, ベジェ曲線を入力画像に重ねる際に, ランダム視点変換を加える. これにより, 人間が落書きを再現する際に起きるずれに対して頑健な攻撃を生成できる. また, 私達は落書きが占める面積に対して L1 正則化を加える. これにより, コンパクトな形状でありながらも, モデルの誤分類を効果的に引き起こす落書きを生成できる.

3 実験

3.1 人間による攻撃の再現

私達は, 人間によって複製された敵対的落書きの有効性を評価する. ImageNet 1K に基づく 10 クラス分類タスクを設定し,CLIP ViT-B/32 モデルを攻撃する. 選んだクラスは, fish, bird, cat, turtle, elephant, spider, crab, dog, fox, butterfly の 10 クラスである. 私達は, ImageNet 1K から各クラスごとにランダムに 1 つの画像サンプルを選択する. 攻撃を加える前は, 全て画像が正しく分類されることが確認された.

それぞれの画像について, 残りの 9 つのクラスとの組み合わせを考え, 合計 $9 \times 10 = 90$ のケースを用意する. 次に, 全てのケースに対して最適化を行い, 敵対的落書きを生成する. 次に, 私達はコンピュータ上で最適化された落書きが CLIP 分類器を欺くことに成功した 78 サンプルを収集し, 4 つのサブセットに分割する. 次に, 本学学生から 4 人の被験者を募り, それぞれに 1 つのサブセットを与える. 私達は 4 人の被験者に, 生成された敵対的な落書きをタブレットを用いて複製してもらう. 被験者にサブセットを与える際には, 私達はサブセット内の各ケースのターゲットクラスに関する情報を隠す. 最後に, 人間が再現した敵対的落書きが CLIP モデルを欺くかどうかを評価する.

実験の結果, 人間の被験者によって再現された攻撃は, 78 件中 44 件成功した (56%). 生成された落書きの形は, スケッチの形をしたものと, 文字列の形をしたものが観察された.

Interpretable and Human-drawable Adversarial Attacks
†Ryoya Nara, The University of Tokyo
‡Yusuke Matsui, The University of Tokyo

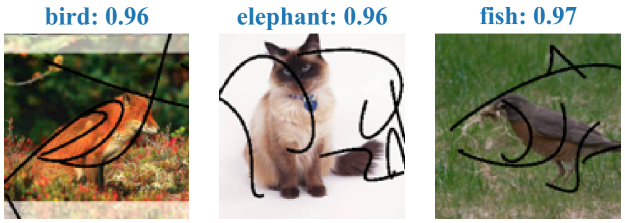


図 1: スケッチの形をした攻撃の例.

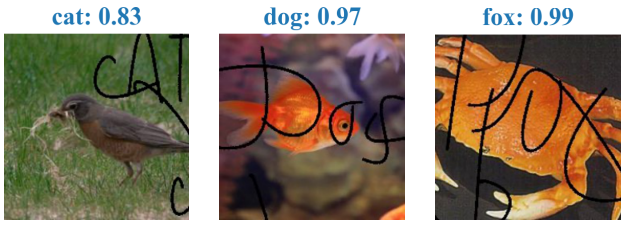
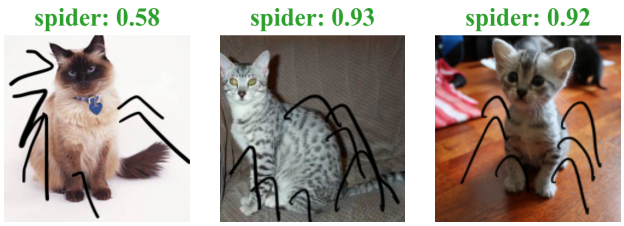


図 2: 文字の形をした攻撃の例.

(a) bird→butterfly



(b) cat→spider

図 3: 記述可能な洞察の例. 図 3a の場合も図 3b の場合も，一番左のケースが人間によって落書きが再現されたケースで，そこから洞察を得て別の画像に落書きを加えたものが真ん中と右のケースである.

スケッチの形をした攻撃

図 1 に示すように，ターゲットクラスに似た落書きが生成されるケースが多く見られた．図 1 の一番左の例だと，キツネが鳥に似るように落書きが生成されている.

文字の形をした攻撃

また, 図 2 に示すように, 曲線の集合がターゲットクラスを表す文字列の形状を持つケースも多く観察された. この結果は，CLIP のタイポグラフィック・バイアス [4] と強く相関している．CLIP は画像内のテキストの存在に強く影響される．例えば, 白い紙に “iPod” と書いてリンゴの上に貼り付ければ, CLIP モデルはこの物体を iPod と分類する．タイポグラフィック・バイアスはすでに記述可能な洞察として知られているが, 敵対的な例を生成することで発見したのは私達が初めてである.

3.2 記述可能な洞察

私達は，生成された敵対的落書きを解釈し, 洞察を得る. その後, 私達はその洞察を用いて敵対的落書きを描き, その落書きが CLIP 分類器を欺くかどうかを評価する．CLIP 分類器に対する記述可能な洞察として, タイポグラフィック・バイアス [4] が既に知られているため, 私達はスケッチの形をした攻撃から洞察を得ることに焦点を当てる.

私達は実験の結果，いくつかのスケッチの形をした敵対的な落書きは, CLIP モデルの画像認識の仕方についての記述可能な洞察を与えることを発見した．図 3a は,“鳥の画像に, 頭に 2 本の線, 体に三角形, 三角形の内側に 2 本の線を追加すると, CLIP はその画像を蝶として誤分類する” という発見である． 図 3b は,“猫の画像にクモの足のようなストロークを追加すると, CLIP はその画像をクモとして誤分類する” という発見である.

4 まとめ

本論文では, 解釈可能で人間が描画可能な形状を持つ敵対的落書きと呼ばれる新しいタイプの攻撃を提案した．私達は, ランダム視点変換と落書き領域を加えながら bézier 曲線の集合の制御点を最適化することで，敵対的落書きを生成した．私達は, 敵対的な落書きが人間によって再現された場合でも CLIP 分類器を欺くことを実験的に示した．さらに私達は, 落書きの形と分類器の出力との間に記述可能な洞察を見出した.

参考文献

[1] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.
[2] S. Casper, M. Nadeau, D. Hadfield-Menell, and G. Kreiman. Robust feature-level adversaries are interpretability tools. In *NeurIPS*, 2022.
[3] T.-M. Li, M. Lukáč, M. Gharbi, and J. Ragan-Kelley. Differentiable vector graphics rasterization for editing and learning. *ACM Trans. Graph.*, 11 2020.
[4] G. Goh, N. Cammarata, C. Voss, S. Carter, M. Petrov, L. Schubert, A. Radford, and C. Olah. Multimodal neurons in artificial neural networks. *Distill*, Vol. 6, No. 3, 2021.