漢字形状敵対的パッチによる物体検出回避の提案

冨岡 佑斗 †　　菊池 浩明 †

明治大学総合数理学部 †

1　はじめに

深層学習は画像認識分野で高い精度を達成する一方で, 敵対的攻撃（Adversarial Attacks）という脆弱性が存在することが知られている [1]. 特に, 実在のオブジェクトに意図的なパターン（Patch）を貼り付けて誤認識を誘発させる敵対的パッチ攻撃 [2] は, 監視カメラや自動運転システムに対する現実世界での脅威として懸念されている. Thys らは [2] で, 精巧に設定されたパッチを印刷して貼り付け人物検出を回避できることを示した. しかし, 生成されるパッチは極彩色のノイズパターンであり, 不自然で攻撃の意図が容易に検出されるという課題が残る.

そこで, 本研究では, 人工的なパッチの代わりに, 日常生活に自然に存在する漢字に着目する. 漢字ならば, 視覚的な自然さ（可読性）を維持したまま, 物体検出モデルに対する検出回避攻撃を実現することが期待できるからである. 従来の敵対的損失関数を基にして, 生成パッチと元の漢字画像との平均二乗誤差（MSE）を導入した手法を提案する. 本手法の導入により, パッチの可読性を大幅に向上させ, カモフラージュ性の高い攻撃を試みる.

本稿では, YOLOv2 モデル [3] を対象に提案手法の評価実験を報告する. 従来のノイズベースの手法の攻撃成功率（ASR）62.03% に対し, 提案手法は 44.06% の ASR を達成した. これにより, 人間には漢字として認識される自然さを保ちつつ, AI モデルに対しては一定の攻撃効果を持つことを明らかにした.

2　提案手法

2.1　システム概要と攻撃対象

本研究では, 漢字画像を初期値とし, 検出回避と形状維持を両立するパッチを生成する. 攻撃物体検出モデルには YOLOv2[3] を対象にする. 図 1 に, 従来手法（2）と提案手法（3）との比較概要を示す. 学習システムの実装

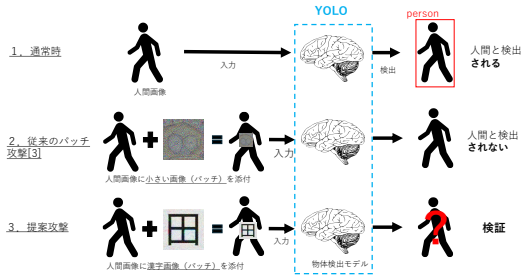

図 1　提案手法の概要と従来手法との比較

は Thys ら [2] のソースコード*をベースとした. 計算機環境にはオンプレの GPU サーバ (NVIDIA RTX A6000 48GB × 2) を使用した.

2.2　損失関数の設計

学習プロセスでは, 以下の損失関数を最小化するようにパッチの画素値を更新する.

Thys ら [2] の損失関数 (Baseline) は

$$L_{Baseline} = L_{obj} + \alpha L_{nps} + \beta L_{tv} \tag{1}$$

と定めている. ここで, L_{obj} は Objectness（バウンディングボックス内に物体が存在する確信度）であり, これを最小化することで非検出を目指す. L_{nps} (Non-Printability Score) は印刷不可能な色を抑制する項, L_{tv} (Total Variation) は画像の滑らかさを保つ項である. α と β はそれぞれの重みを調整する係数である.

本研究では, 漢字の形状維持を最優先するため, L_{nps} と L_{tv} の代わりに, 形状維持のための損失 $L_{readability}$ を導入した新たな損失関数

$$L_{proposed} = L_{obj} + \gamma L_{readability} \tag{2}$$

を提案する. ここで $L_{readability}$ は, 元の漢字画像 P_{org} と生成されたパッチ P_{adv} との平均二乗誤差（MSE）

$$L_{readability} = \frac{1}{N} \sum (P_{adv} - P_{org})^2$$

として定義される. γ は攻撃力と形状維持のバランスを調整するハイパーパラメータで, N はパッチの大きさ (W × I) のピクセル数である.

†Yuto Tomioka and Hiroaki Kikuchi, Department of Frontier Media Science, School of Interdisciplinary Mathematical Science, Meiji University.

*https://gitlab.com/EAVISE/adversarial-yolo

表 1　各手法における攻撃性能の比較

手法	Recall	Precision	MP	ASR
Clean	0.971	0.638	0.923	0.00
Random	0.918	0.594	0.862	0.055
Baseline	0.369	0.301	0.231	0.620
Baseline+MSE	0.529	0.323	0.348	0.455
Proposed	0.543	0.355	0.355	0.441

2.3　評価指標

攻撃の有効性評価には, 奥田ら [4] を参考に, 平均適合率 Mean Precision(MP), 攻撃成功率 (ASR) を用いる. ASR は, クリーンデータで正しく検出されていた物体が, 攻撃によって検出失敗した割合を示す.

2.4　実験結果

本実験では, 提案手法の有効性を検証するため, 以下の 3 つの条件で敵対的パッチの生成を行った.

- **Baseline**：[2] と同様の損失関数式 (1) を用いる.
- **Baseline+MSE**：Baseline に形状維持損失を加えた $L_{baseline+MSE} = L_{obj} + \gamma L_{readability} + \alpha L_{nps} + \beta L_{tv}$.
- **proposed**：提案手法. 形状維持を最優先し, 滑らかさや印刷可能性の制約を除く式 (2).

全ての条件において, パッチの初期画像には漢字の「田」を使用し, 学習は 1000 エポックで統一した. 漢字として「田」を選定した理由は, 直線で構成された単純な幾何学形状であり, 黒画素の密度が高く攻撃に必要な摂動を埋め込みやすいと考えたためである.

2.5　定量的評価

表 1 に, 各手法における評価結果を示す. 提案手法 (proposed) の ASR は 44.06% であり, Baseline(62%) には及ばないものの, Random（5.51%）と比較して著しく高い値を示した. これは, 漢字の形状を維持した状態でも, YOLOv2 モデルに対して十分な攻撃が可能であることを示している.

図 2 に学習後の生成パッチを示す. Baseline（図 2(b)）は元の「田」の原型を留めておらず, 人間には意味不明な模様に見える. 対して, Baseline+MSE 手法（図 2(c)）及び提案手法（図 2(d)）は, 色彩の変化やノイズを含みつつも,「田」という文字形状を明瞭に維持している. また, (c) と (d) を比較すると, 滑らかさ（L_{tv}）の制約がない proposed 手法 (d) の方が, エッジが鋭く元の漢字の原型をより強く留めている.

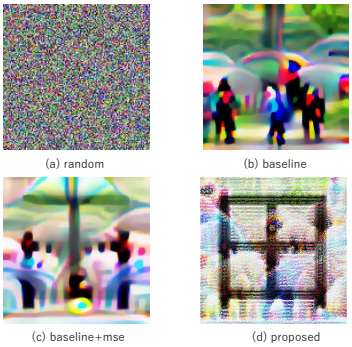

図 2　学習により生成されたパッチの比較

s

3　まとめ

本研究では, 漢字形状を維持した敵対的パッチの生成手法を提案し, 検証した. 実験の結果, 提案手法は, 従来のノイズベースの手法と比較して攻撃性能は低下するものの, 視覚的な自然さ（可読性）を向上させるていた.

今後は対象とする漢字の種類を増やし, 画数や形状の違いが攻撃性能に与える影響を詳細に分析する.

参考文献

[1] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli: ”Evasion attacks against machine learning at test time”, *Proc. ECML PKDD*, pp. 387–402 (2013).

[2] Thys, S., Van Ranst, W., and Goedemé, T.: ”Fooling automated surveillance cameras: adversarial patches to attack person detection”, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, pp. 49-55 (2019).

[3] Redmon, J. and Farhadi, A.: ”YOLO9000: Better, Faster, Stronger”, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 7263-7271 (2017).

[4] 奥田将登, 吉田康太, 藤野毅：「物体検出器 YOLOv8 に対する敵対的パッチ生成攻撃とマルチスペクトル画像を用いた緩和手法の評価」, 暗号と情報セキュリティシンポジウム (SCIS), 1-3G3-4 (2025).