

ViT ベースモデルのための視覚情報保護を考慮した物体検出法

A privacy-aware object detection method for ViT-based models

末吉保稀[†]

西川清史[‡]

貴家仁志[‡]

[†]東京都立大学

Homare Sueyoshi[†]

Kiyoshi Nishikawa[‡]

Hitoshi Kiya[‡]

[†]Tokyo Metropolitan University

アブストラクト テスト画像の視覚情報保護を考慮した物体検出法を提案する。視覚情報保護に関する先行研究は、画像分類タスクが中心である。本稿では、知覚暗号化を用いた物体検出法を初めて提案する。提案法は、Vision Transformer (ViT) の埋め込み構造を利用して、非暗号化時と同等の精度を有するモデルを構築することができる。実験では、ViT に基づく物体検出モデルある ViTDeT を例にして、提案法の有効性を確認する。

1 はじめに

深層学習モデルの学習には膨大なデータと計算コストが必要とされることから、クラウドサーバーを使用することが多い。さらに、学習済みモデルをクラウドサーバ上に置き、サービスを展開することが期待されている。しかし、画像に代表されるデータは、一般に多くの個人情報を含む。一方、クラウド環境は信頼できるとは言えず、クラウドサーバーを安全に使用するにはデータのプライバシー保護を考慮する必要がある。本稿では、このような背景から、Vision Transformer (ViT) [1] を用いた ViTDeT [2] と呼ばれるモデルを例にして、プライバシー保護を考慮した物体検出の新しい手法を提案する。提案手法では、モデルとテスト画像の両方に対して暗号化を施し、画像の視覚情報保護を目指す。

2 提案手法

2.1 概要

実行手順の要約は以下の通りである。

- 1) モデル作成者は、学習済みモデル f_θ 内のパッチ埋め込み E を秘密鍵 k を用いて \hat{E} に変換し、その暗号化されたモデル \hat{f}_θ をクラウド上にアップロードする。
- 2) ユーザーはモデル作成者から秘密鍵 k を受け取り、テスト画像に対して暗号化を行い、クラウド上にアップロードする。
- 3) 暗号化されたテスト用画像に対して、 \hat{f}_θ によって物体検出が実行され、テスト結果が出力される。

上記の枠組みにおいて、クラウドサーバーは、視覚情報が保護されたテスト画像を受け取り、かつ鍵に関する情報を持たない。

2.2 脅威モデル

攻撃者の目的は、暗号化画像から視覚情報を復元し、個人情報を獲得することである。攻撃者は、暗号化画像及び \hat{f}_θ にアクセスでき、かつ暗号化アルゴリズムを既知とするが、秘密鍵を持たないと仮定する。知覚暗号化法と攻撃耐性の関係は、種々の攻撃法に対して評価されている [3, 4]。上述の条件から、一般に暗号文単独攻撃 (cipher-text-only attack (COA)) の下で、知覚暗号化の攻撃耐性は評価される。

2.3 ランダム置換行列の作成方法

- 1) 秘密鍵 k を用いて、長さ $L = p^2c$ のランダム整数ベクトル $l = [l(1), l(2), l(i), \dots, l(L)]$ を生成する。 p はパッチサイズ、 c は画像のチャンネル数とする。ここで、 l の要素 $l(i)$ は、 $l(i) \in \{1, 2, \dots, L\}$ および $i \neq j \Rightarrow l(i) \neq l(j)$, $i, j \in \{1, 2, \dots, L\}$ を満たす。
- 2) ランダム置換行列 $E_{\text{enc}} \in \mathbb{R}^{L \times L}$ の要素 $k(i, j)$ を $k(i, j) = 0$ if $l(i) \neq j$, 1 if $l(i) = j$ ように定義する。 E_{enc} は直交行列となり、 $E_{\text{enc}}^{-1} = E_{\text{enc}}^\top$ が成立する。

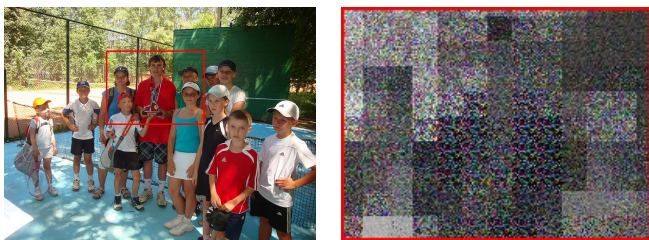
2.4 モデル暗号化

E_{enc} を用いて、学習済みモデルの E を \hat{E} に変換する。すなわち、 $\hat{E}^\top = E_{\text{enc}} E^\top$, $\hat{E} \in \mathbb{R}^{L \times D}$ とする。

2.5 画像暗号化

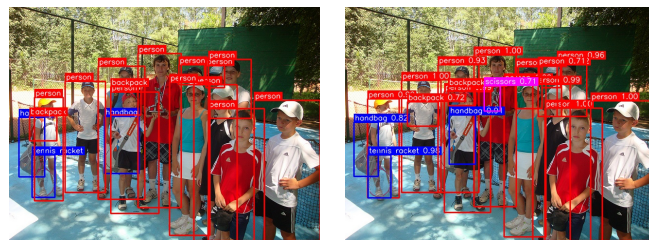
テスト画像 $x \in \mathbb{R}^{h \times w \times c}$ を、大きさ $p \times p$ の N 個の非重複ブロック $B = [B_1, \dots, B_i, \dots, B_N]$, $B_i \in \mathbb{R}^{p \times p \times c}$ に分割し、各ブロックをベクトル $b_i \in \mathbb{R}^L$ として平滑化する。 $E_{\text{enc}} \in \mathbb{R}^{L \times L}$ を用いて、 $\hat{b}_i = b_i E_{\text{enc}}^\top$ のように画素値を置換する。

次に、各ベクトル \hat{b}_i を暗号化ブロック \hat{B}_i に戻し、それらを連結させて暗号化画像 \hat{x} を構築する。図 1 に $p = 16$ とした時の暗号化画像の例を示す。



非暗号化画像 (原画像) 暗号化画像 (拡大画像)

図 1: 非暗号化画像と暗号化画像の比較



正解ラベル 提案法 (mAP=49.327)

図 2: 物体検出結果 (ViT-B)

表 1: 物体検出の精度比較

backbone	method	mAP	mAP_s	mAP_m	mAP_l
ViT-B	Baseline	49.329	32.160	52.935	65.470
	Proposed	49.327	32.146	52.920	65.478
ViT-L	Baseline	54.050	36.555	58.598	68.939
	Proposed	54.050	36.502	58.604	68.938

暗号化画像は、暗号化されたモデル \hat{f}_θ へ入力される。その際、 E_{enc} の影響は \hat{E} と暗号化画像との対の関係によって打ち消される。

3 実験

3.1 実験準備

本実験においては、COCO データセット [5] 内の Val2017 と呼ばれる 5000 枚の検証用データに対してテストを行った。モデルには、公式 Github 上のモデルを用いた。そのモデルは、ViT-B, ViT-L に基づいた “mae_pretrain_vit_base.pth” および、 “mae_pretrain_vit_large.pth” を事前学習モデルとして用いてファインチューニングを行ったものである。精度を表す指標として、平均適合率 (mAP) を用いた。これは 0 以上 100 以下の値を取り、数値が大きいほど精度が良いことを示している。また、mAP_s, mAP_m, mAP_l はそれぞれ小物体、中物体、大物体に対する mAP の値を示している。

3.2 実験結果

表 1 は、種々の条件下で推定されたバウンディングボックスの精度の評価結果である。Baseline はモデルおよび画像共に非暗号化時の結果である。図 2 に、ViT-B を初期の重みとし、閾値を 0.7 とした場合の物体検出例を示す。表 1 より、提案法は、非暗号化時と同等の物体検出精度を持つことが確認できた。

4 まとめ

本稿では、視覚的情報を保護しながら物体検出を実行可能とする、ViT の埋め込み構造を利用した方法を提案

した。ViTDeT モデルを例にして、提案法は、視覚情報を保護できるだけでなく、非暗号化と同等の精度が得られることが確認された。

今後の課題として、視覚情報の保護能力の制御と、各種攻撃に対する防御性能を評価する必要がある。

謝辞

本研究の一部は、JSPS 科研費 25K07750 の助成を受けた。

参考文献

- [1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [2] Y. Li, H. Mao, R. Girshick, and K. He, “Exploring Plain Vision Transformer Backbones for Object Detection,” *arXiv preprint arXiv:2203.16527*, 2022.
- [3] H. Kiya, A. P. M. Maung, Y. Kinoshita, S. Imaizumi, S. Shiota, et al., “An overview of compressible and learnable image transformation with secret key and its applications,” *APSIPA Transactions on Signal and Information Processing*, Vol. 11, No. 1, 2022.
- [4] A. P. Maung Maung, I. Echizen, and H. Kiya, “On the Security of Learnable Image Encryption for Privacy-Preserving Deep Learning,” *IEEE Access*, vol. 12, pp. 126415-126425, 2024.
- [5] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, “Microsoft COCO: Common Objects in Context,” *arXiv preprint arXiv:1405.0312*, 2015.
- [6] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, “Detectron2,” *GitHub repository*, 2019.