SCIS 2022 2022 Symposium on
Cryptography and Information Security
Osaka, Japan & Online, Jan. 18 – 21, 2022
The Institute of Electronics,
Information and Communication Engineers

物体検出 CNN に対する複数配置に着目した遠隔 Adversarial Patch 攻撃
Multiple Adversarial Patches on Object Detection using CNN

大西　健斗 *　　中井　綱人 *　　鈴木　大輔 *
Kento Oonishi　　Tsunato Nakai　　Daisuke Suzuki

キーワード CNN，YOLO，Adversarial patch

あらまし

本発表では，YOLOv2 [3] に対し，複数の遠隔 Adversarial patch を配置する攻撃方法の提案を行う．YOLOv2 は，一度の処理で対象物の分類が可能な CNN であり，CNN に対しては，誤検知を引き起こす攻撃方法等，現在まで，さまざまな攻撃方法が提案されている．CNN に対する強大な脅威の一つとして，敵対的な特徴を持つ patch を利用した攻撃方法がある [1, 4]．この攻撃方法では，事前に用意された patch を配置することで，誤検知を引き起こすことが可能である．特に，この攻撃方法は，patch を配置すれば攻撃が可能であるため，実現可能性が高く，YOLOv2 のような物体識別システムに対する現実的な脅威となりうる．したがって，patch 攻撃の脅威を見積もることは，安全な CNN の利用法を考察するために必要不可欠である．

本稿では，Saha らの遠隔 patch 攻撃の実装 [4] に基づき，YOLOv2 の CNN に大きな影響を与える patch に関する理論的解析を行った．特に，図 1 を含む公開データセットである PASCAL VOC データセット [2] の各画像に patch を配置し，その効果を検証した．まず，YOLOv2 が，畳み込みとプーリングで構成されている点に着目し，patch 情報の拡散がどのように起こるか，について理論的解析を行った．特に，本稿では，patch の面積が同一となるような配置について解析を行った．解析の結果，単一遠隔 patch とは異なり，図 1 のように配置された複数遠隔 patch の場合，図 2 のように patch の影響が画面全体に広がり，patch が YOLOv2 の CNN に与える影響が大きいことを示した．さらに，遠隔 patch が 2 枚の場合，表 1 の通り，対象クラスの検出率が減少することを実験により示した．

表 1: 同面積の Patch における検出率の変化

枚数及び面積	配置方法	bicycle	bus	car	person
単一	左上	66.21%	56.38%	53.40%	55.55%
100×100 =	左下	47.22%	43.85%	41.09%	59.90%
10,000 pixel	左中央	31.45%	22.71%	31.51%	49.85%
複数 (2 枚) 70×70×2 = 9,800 pixel	左右中央 (図 1)	6.30%	0.22%	8.19%	19.89%

図 1: patch 2 枚の配置図

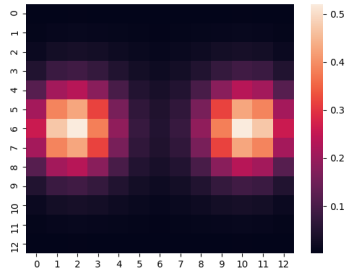

図 2: 図 1 における patch 2 枚の影響範囲

参考文献

[1] Brown, T. B., Mané, D., Roy, A., Abadi, M., and Gilmer, J., “Adversarial Patch.” eprint arXiv 1712.09665 (2017).

[2] Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A., “The Pascal Visual Object Classes Challenge: A Retrospective.” International Journal of Computer Vision, 111(1), 98–136 (2015).

[3] Redmon, J. and Farhadi, A., “YOLO9000: Better, Faster, Stronger.” IEEE CVPR 2017, pp. 6517–6525 (2017).

[4] Saha, A., Subramanya, A., Patil, K., and Pirsiavash, H., “Role of Spatial Context in Adversarial Robustness for Object Detection.” IEEE/CVF CVPRW 2020, pp. 3403–3412 (2020).

* 三菱電機株式会社情報技術総合研究所，〒 247-8501 神奈川県鎌倉市大船 5-1-1 Mitsubishi Electric, 5-1-1 Ofuna, Kamakura-shi, Kanagawa, 247-8501 Onishi.Kento@ap.MitsubishiElectric.co.jp