Adversarial Attack and Defense of YOLO Detectors in Autonomous Driving Scenarios

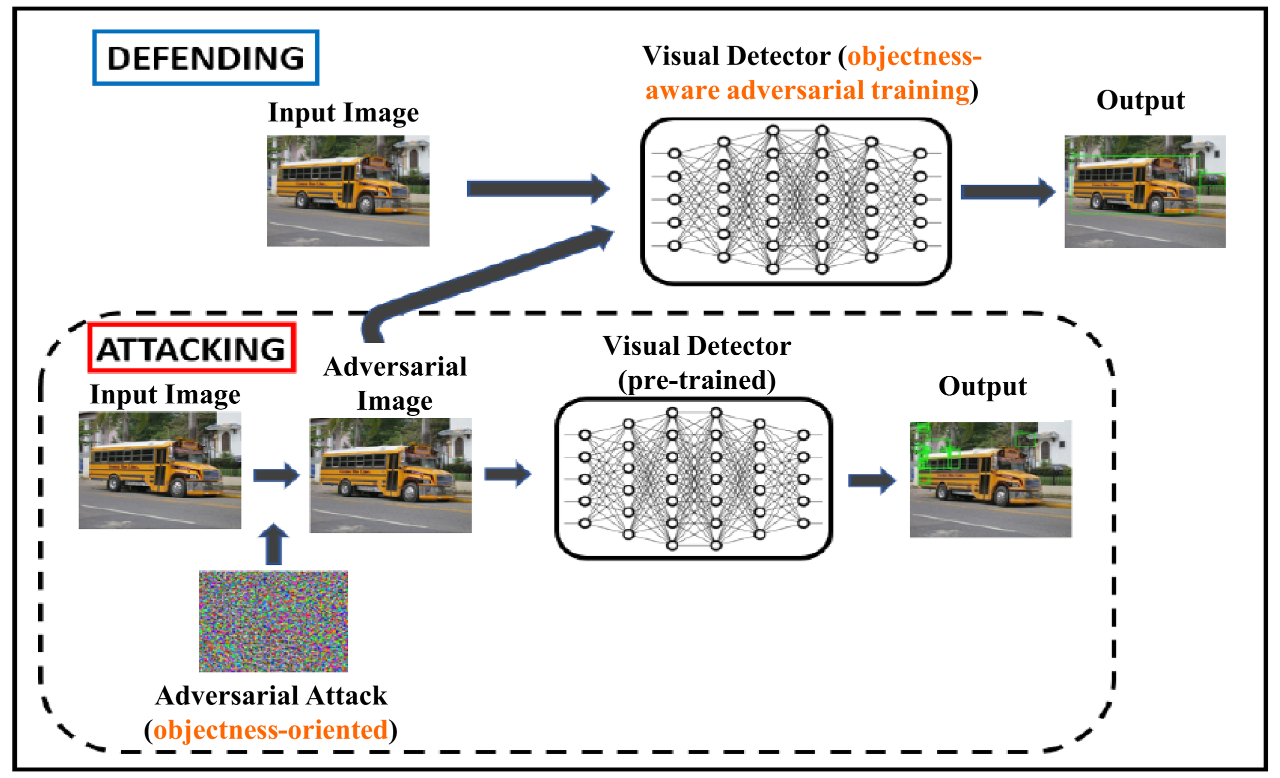
Jung Im Choi and Qing Tian

Department of Computer Science, Bowling Green State University

IV22 33rd IEEE INTELLIGENT VEHICLES SYMPOSIUM

Overview & Research Goals

- Deep visual detectors are **vulnerable to adversarial attacks**. A comprehensive understanding of YOLO detectors' vulnerability is needed before their robustness can be improved. However, only a few adversarial attack/defense works have **focused on visual detection, especially in autonomous driving**.
- **Research Goals:**
  - **Goal 1 (Attacking):** To understand the adversarial vulnerability of deep visual detectors and design **effective adversarial attacks** by utilizing them
  - **Goal 2 (Defending):** To improve detectors' robustness via a new **adversarial training-based approach**



Contributions

- We identified a serious vulnerability in YOLOs which comes from the **objectness aspect** and proposed a more effective objectness-oriented adversarial attack approach.
- We found the direction of the image gradient derived from the objectness loss is more consistent with those from the classification and localization losses.
- We proposed a new defense strategy explicitly paying attention to the objectness aspect.
- Our objectness-aware adversarial training framework can help alleviate the potential conflicts/misalignment of the image gradients sourced from different loss components.

Methods

Decomposition of Adversarial Vulnerability in YOLO

- Overall loss in YOLO consists of three components (i.e., objectness, localization, & classification losses):

$$L(\mathbf{x}, y, \mathbf{b}; \theta) = L_{OBJ}(\mathbf{x}, \mathbf{b}; \theta) + L_{LOC}(\mathbf{x}, \mathbf{b}; \theta) + L_{CLS}(\mathbf{x}, y; \theta)$$

Objectness-Oriented Adversarial Attack

- Unlike prior works, we **explicitly leverage objectness loss** in addition to localization and classification losses to generate adversarial examples for visual detection in self-driving scenarios:

$$\mathbf{x}'_{obj,\,PGD} = \mathcal{P}(\mathbf{x} + \alpha \cdot sign(\nabla_{\mathbf{x}} L_{OBJ}(\mathbf{x}, \mathbf{b}; \theta)))$$
$$\mathbf{x}'_{loc,\,PGD} = \mathcal{P}(\mathbf{x} + \alpha \cdot sign(\nabla_{\mathbf{x}} L_{LOC}(\mathbf{x}, \mathbf{b}; \theta)))$$
$$\mathbf{x}'_{cls,\,PGD} = \mathcal{P}(\mathbf{x} + \alpha \cdot sign(\nabla_{\mathbf{x}} L_{CLS}(\mathbf{x}, y; \theta)))$$

Objectness-Aware (OA) Adversarial Training

Algorithm 1 Objectness-Aware Adversarial Training

Input: Dataset D, Training epochs N, Batch size B, Perturbation bounds ϵ

Output: Learned model parameter θ

for epoch =1 **to** N **do**
  for random batch $\{\mathbf{x}^i, \{yi, \mathbf{b}^i\}\}_{i=1}^{B} \sim D$ **do**
    $(\mathbf{x}'_{obj})^i = P(\mathbf{x}^i + \epsilon \bullet sign(\nabla_{\mathbf{x}} L_{OBJ}(\mathbf{x}^i, \mathbf{b}^i; \theta)))$
    $(\mathbf{x}'_{loc})^i = P(\mathbf{x}^i + \epsilon \bullet sign(\nabla_{\mathbf{x}} L_{LOC}(\mathbf{x}^i, \mathbf{b}^i; \theta)))$
    $(\mathbf{x}'_{cls})^i = P(\mathbf{x}^i + \epsilon \bullet sign(\nabla_{\mathbf{x}} L_{CLS}(\mathbf{x}^i, y^i; \theta)))$
    Choose $\underline{\mathbf{x}}^i$ that leads to the max total loss:
    $\underline{\mathbf{x}}^i = \arg\max_{\tilde{x}^i \in \{(x'_{obj})^i, (x'_{loc})^i, (x'_{cls})^i\}} L(\tilde{x}^i, \{y^i, \mathbf{b}^i\}; \theta)$
    Perform an adversarial training step w.r.t. θ:
    $\arg\min_{\theta} L(x^i, \{y^i, \mathbf{b}^i\}; \theta) + L(\underline{\mathbf{x}}^i, \{y^i, \mathbf{b}^i\}; \theta)$
  end for
end for

Experiments & Results

Datasets

	KITTI	COCO_traffic
Classes #	3	8
Train. Img. #	3,740	71,536
Test Img. #	3,741	3,028

Attack Design:

- Used a range of different attack sizes ϵ = [2, 4, 6, 8]
- Adapted both FGSM and PGD using L_∞ norm

Figure 1. Distribution of adversarial examples derived from the three task losses in YOLO

Qualitative Analysis of Attacks on KITTI and COCO_traffic

Figure 2. Visual comparison of detection results under different task-specific attacks (top: KITTI, bottom: COCO_traffic).

Results – cont.

Quantitative Analysis of Attacks

- The objectness-oriented attacks ($\mathcal{A}_{obj}$) are more effective than $\mathcal{A}_{loc}$ and/or $\mathcal{A}_{cls}$.

Method	Att. Size	$\mathcal{A}_{loc}$	$\mathcal{A}_{cls}$	$\mathcal{A}_{loc+cls+obj}$	$\mathcal{A}_{obj}$
PGD-10 (KITTI)	ϵ = 2	-1.22	-0.87	-42.44	**-42.64**
	ϵ = 4	-4.11	-2.64	-51.47	**-51.67**
	ϵ = 6	-7.00	-5.91	-57.17	**-54.39**
	ϵ = 8	-10.66	-9.59	-55.48	**-55.83**
PGD-10 (COCO)	ϵ = 2	-0.19	-0.15	-36.55	**-37.55**
	ϵ = 4	-0.70	-0.77	-43.84	**-43.93**
	ϵ = 6	-1.88	-2.26	-45.31	**-45.69**
	ϵ = 8	-3.24	-3.58	-46.88	**-47.08**

Table 1. Performance degradation comparison. $\mathcal{A}_{loc}$, $\mathcal{A}_{cls}$, $\mathcal{A}_{obj+loc+cls}$, & $\mathcal{A}_{obj}$: attacks sourced from corresponding task losses.

Adversarial Training Results

- The models adversarially trained with objectness-based attacks ($\mathcal{M}_{OBJ}$ and $\mathcal{M}_{OA}$) leads to more robustness than those utilizing other two task losses.

KITTI

Model	$\mathcal{A}_{obj}$	$\mathcal{A}_{loc+cls+obj}$
$\mathcal{M}_{STD}$	28.43	28.63
$\mathcal{M}_{ALL}$	39.65	40.65
$\mathcal{M}_{MTD}$	36.13	35.94
$\mathcal{M}_{LOC}$	37.86	37.61
$\mathcal{M}_{CLS}$	39.29	39.70
$\mathcal{M}_{OBJ}$	**49.43**	**48.83**
$\mathcal{M}_{OA}$	42.26	41.86

COCO_traffic

Model	$\mathcal{A}_{cls}$	$\mathcal{A}_{loc+cls+obj}$
$\mathcal{M}_{STD}$	22.17	22.29
$\mathcal{M}_{ALL}$	34.58	33.44
$\mathcal{M}_{MTD}$	33.26	33.20
$\mathcal{M}_{LOC}$	33.23	32.10
$\mathcal{M}_{CLS}$	31.71	31.58
$\mathcal{M}_{OBJ}$	33.30	32.69
$\mathcal{M}_{OA}$	**34.77**	**33.61**

Table 2. Performance comparison of **adversarially trained YOLO models.** $\mathcal{M}_{STD}$: trained standardly, $\mathcal{M}_{MTD}$: trained using the multi-task domain algorithm [1], $\mathcal{M}_{OA}$: trained using our OA algorithm.

Conclusion

- We identified a serious vulnerability of YOLO detectors in autonomous driving scenarios.
- We proposed: (1) an effective attack strategy targeting the objectness loss in visual detection, and (2) an objectness-aware adversarial training framework.
- Experiments on both datasets showed the effectiveness of our proposed approaches.

Reference - [1] H. Zhang and J. Wang, "Towards adversarially robust object detection," ICCV'2019