Evaluating the Adversarial Robustness of Detection Transformers

Amirhossein Nazeri*, Chunheng Zhao*, Prof. Pierluigi Pisu
anazeri@clemson.edu, chunhez@clemson.edu, pisup@clemson.edu
*Equal Contribution

Abstract

This research presents a comprehensive evaluation of **DETR** model and its variants under **white-box and black-box adversarial attacks**, using the **MS-COCO and KITTI** datasets to cover general and autonomous driving scenarios. We extend prominent white-box attack methods (FGSM, PGD, and C&W) to assess DETR's vulnerability, demonstrating that DETR models are significantly susceptible to adversarial attacks, like traditional CNN-based detectors. Our extensive transferability analysis reveals **high intra-network** transferability among DETR variants but **limited cross-network** transferability to CNN-based models. Additionally, we propose a novel untargeted attack designed specifically for DETR, exploiting its intermediate loss functions to induce misclassification with minimal perturbations. *Performed train/evaluation on Clemson Palmetto HPC with Nvidia V100 &A100 GPUs for ~600 hours.*

Motivation

Robust object detection is critical for autonomous driving and mobile robotics, where accurate detection of vehicles, pedestrians, and obstacles is essential for ensuring safety. Despite the advancements in object detection transformers (**DETRs**), their robustness against adversarial attacks remains **underexplored, ,** with **no comprehensive evaluation** of their performance under standard **white-box** and **black-box** attacks.

Approach

- *White-box attacks extension*: FGSM, PGD, C&W.
- *Transferability-based black-box attacks*:

$$TR_{m,n} = \frac{AP^m_{clean} - AP^m_{adv(n)}}{AP^n_{clean} - AP^n_{adv(n)}}$$

- *Our white-box attack designed for DETR:*

(1) initialize the attack with a one-step slight perturbation. (FGSM-based)

(2) apply a multi-step C&W process with loss from both final and intermediate layers to the slightly-perturbed images.

Algorithm 1 Our Attack on DETR

Input: initial image x, ground-truth label t, steps m, perturbation constant c, initial perturbation constant α.
Output: adversarial image x_{adv}

Initialize:
$x \leftarrow x + \alpha \cdot \nabla_x(-J_{cls}(\theta, x, t_c))$
$w_0 = zeros(x)$
for $i = 0$ **to** $m - 1$ **do**
  $x_{adv} = \frac{1}{2} \cdot (tanh(w_i) + 1)$
  $Loss_{dm} = L_2(x_{adv}, x)$
  $Loss_{cls} = c \cdot f(x_{adv})$
  $Loss_{bb}^{\{o,k\}} = -J_{bb}^{\{o,k\}}(\theta, x_{adv}, t_a)$
  $Loss_{iou}^{\{o,k\}} = -J_{iou}^{\{o,k\}}(\theta, x_{adv}, t_a)$
  Update w_i with gradient decent
  $w_i \leftarrow \nabla_{w_i}(Loss_{total} = Loss_{dm} + Loss_{cls} + Loss_{bb}^{\{o,k\}} + Loss_{iou}^{\{o,k\}})$
  if $Loss_{total}$ does not converge **then**
    return: $x_{adv} = \frac{1}{2} \cdot (tanh(w_i) + 1)$
  end if
end for
return: $x_{adv} = \frac{1}{2} \cdot (tanh(w_i) + 1)$

Implementation

White-box attacks setup:

Datasets: MS COCO 2017 and KITTI Vision for general and domain-specific tasks. **Model:** DETR variants: DETR-R50, DETR-R50-DC5, DETR-R101, and DETR-R101-DC5. **Attacks:** FGSM: $\varepsilon = 0.03, 0.05, 0.1$. PGD: $\varepsilon = 0.03, 0.1$ 10 iterations. The L_∞ radius: 10/255. C&W: $c = 1, 3, 5$ in 200 iterations. **Ours:** $\alpha = 0.3$, $c = 0.8$ with 200 iterations.
Evaluation metrics: Robustness Score (RS) = AP_{Adv}/AP_{Clean}.

Black-box attacks setup (Transferability analysis):

White-box attacks generated on a surrogate model n and transferred to an unknown target model m. Explored: **Intra-network transferability:** Attack generated on a DETR variant and transferred to another DETR variant. **Cross-network transferability:** Attack generated on a DETR variant and transferred to a traditional CNN-based model. To evaluate intra-network and cross-network transferability on the KITTI dataset, we retrain the four DETR variants as well as a Faster R-CNN.

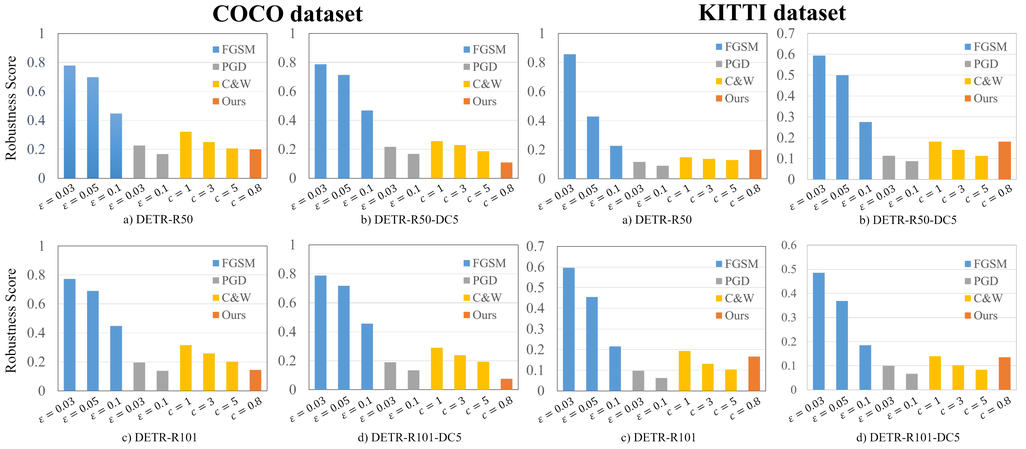

Figure 1. Robustness Score of DETRs on COCO and KITTI.

Figure 2. DETR Object detection results on sample images under various adversarial attacks.

Attacks caused: Misclassification & Mis-localization

Results & Discussion:

- Substantial vulnerabilities in DETR models **even against basic and classic attacks** (FGSM, PGD, C&W), consistent with the vulnerabilities observed in CNN-based object detectors.
- **Strong transferability** of adversarial examples generated by DETR models across different DETR variants, but **limited transferability** to CNN-based object detectors (Faster R-CNN).
- Our attack can cause significant performance degradation with less visible perturbations.

Table. 1 Intra-network and cross-network transferability results on COCO and KITTI dataset.

COCO						
Models	Attacks	Detr-R50	Detr-R50-DC5	Detr-R101	Detr-R101-DC5	Faster R-CNN
Detr-R50	FGSM($\epsilon = 0.05$)	100%	96.8%	71.8%	65.3%	66.9%
	PGD($\epsilon = 0.03$)	100%	109.0%	106.2%	109.3%	54.8%
	C&W($c = 3$)	100%	99.1%	89.5%	80%	66.7%
Detr-R50-DC5	FGSM($\epsilon = 0.05$)	107.5%	100%	78.3%	70.8%	70.0%
	PGD($\epsilon = 0.03$)	101.5%	100%	101.8%	104.7%	51.6%
	C&W($c = 3$)	94.2%	100%	85.5%	87.3%	69.1%
Detr-R101	FGSM($\epsilon = 0.05$)	86.4%	78.8%	100%	90.9%	70.5%
	PGD($\epsilon = 0.03$)	96.8%	100.1%	100%	107.4%	53.3%
	C&W($c = 3$)	82.5%	56.1%	100%	86.9%	59.0%
Detr-R101-DC5	FGSM($\epsilon = 0.05$)	89.7%	82.5%	104.0%	100%	71.4%
	PGD($\epsilon = 0.03$)	93.6%	96.1%	100.3%	100%	51.0%
	C&W($c = 3$)	79.8%	81.0%	91.2%	100%	67.8%
KITTI						
Models	Attacks	Detr-R50	Detr-R50-DC5	Detr-R101	Detr-R101-DC5	Faster R-CNN
Detr-R50	FGSM($\epsilon = 0.05$)	100%	77.9%	73.6%	59.6%	84.1%
	PGD($\epsilon = 0.03$)	100%	92.8%	95.8%	98.8%	77.6%
	C&W($c = 3$)	100%	88.7%	88.1%	87.5%	72.2%
Detr-R50-DC5	FGSM($\epsilon = 0.05$)	116.6%	100%	89.9%	76.3%	105.9%
	PGD($\epsilon = 0.03$)	110.2%	100%	104.2%	108.6%	85.9%
	C&W($c = 3$)	107.3%	100%	99.0%	96.0%	78.8%
Detr-R101	FGSM($\epsilon = 0.05$)	81.8%	62.6%	100%	93.4%	83.8%
	PGD($\epsilon = 0.03$)	101.5%	92.4%	100%	104.8%	83.7%
	C&W($c = 3$)	94.7%	86.3%	100%	100%	76.0%
Detr-R101-DC5	FGSM($\epsilon = 0.05$)	84.5%	69.5%	106%	100%	79.5%
	PGD($\epsilon = 0.03$)	99.7%	89.8%	98.5%	100%	82.6%
	C&W($c = 3$)	93.1%	85.3%	97.3%	100%	72.5%

Acknowledgements

This work is funded by the National Science Foundation CNS No. 2200457.