Benchmarking and deeper analysis of adversarial patch attack on object detectors

Pol Labarbarie (IRT SystemX)

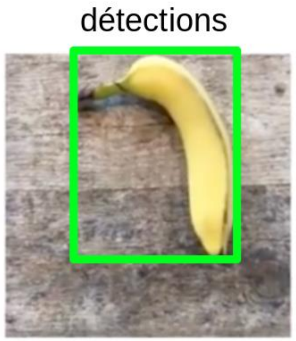
Adrien Chan-Hon-Tong (ONERA)

Stéphane Herbin (ONERA)

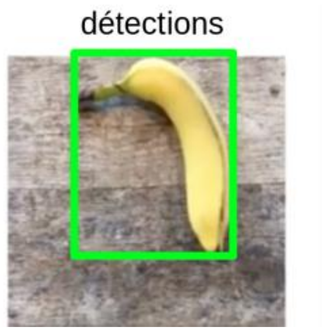
Milad Leyli-Abadi (IRT SystemX)

Classical adversarial attacks

Original image



Adversarial noise



Nothing happened

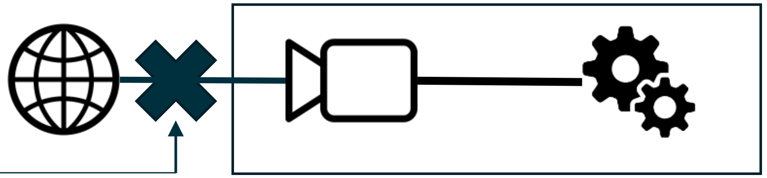
Changed the detected object class

Suppressed the detection

Possible scenarios of a classical adversarial attack on object detector.

Physically feasible?

- Without direct access to sensors



Adversarial noise   Image captured by embedded sensors   détections

: Wrong class detection

: Good detection

Example of a classical adversarial attack perturbating image pixels captured by autonomous vehicle embedded sensors.

How to exactly perturbate image pixels?

Adversarial patch attacks

Differences between classical attacks and patch attacks:

- Unconstrained in magnitude.
- Constrained in space.

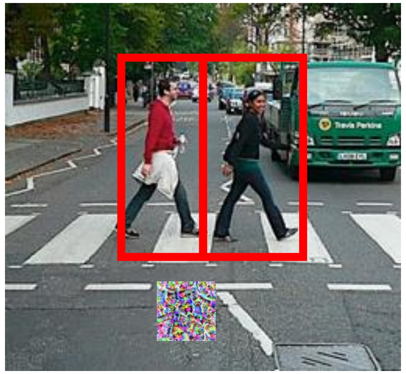
Printed and placed into the scene

Adversarial patch

Image captured by embedded sensors

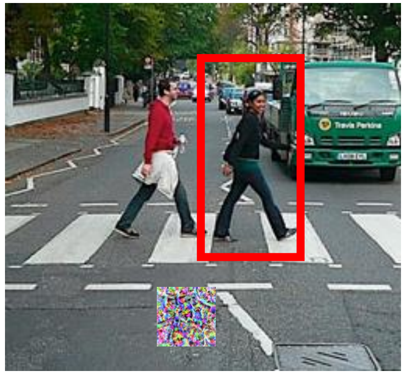
=

détections



Wrong class

détections



Suppress detection

: Wrong class detection

: Good detection

State-of-the-art patch attacks

- Lee *et al.* (Lee *et al.*, 2019): maximizing the YOLO loss over the ground truths.

- Dpatch (Liu *et al.*, 2018): minimizing the YOLO loss but redefined the ground truths boxes at the patch localization.

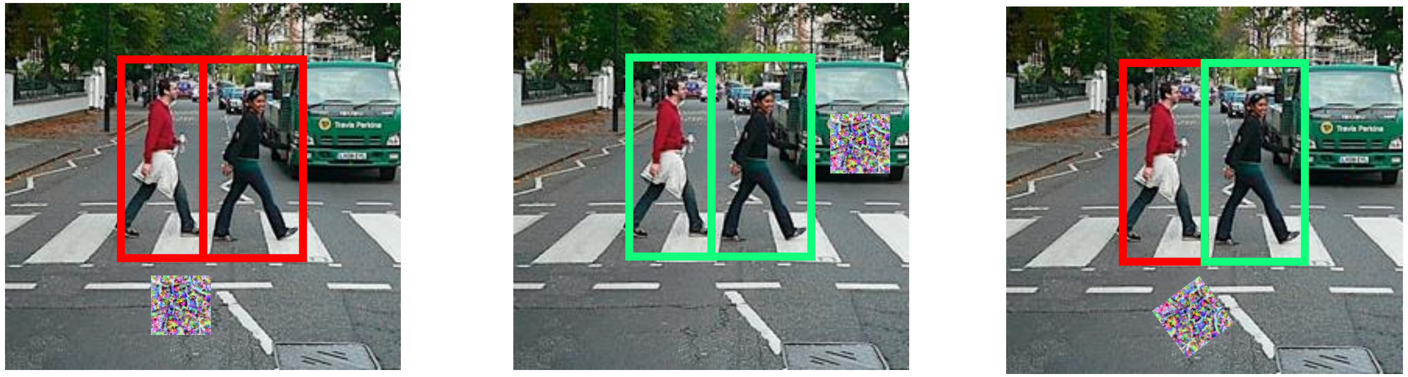
- Saha *et al.* (Saha *et al.*, 2020): minimizing the probability of one chosen class.

Criticality of patch attacks

It seems to be a serious threat to consider

But

How to measure the real criticality of patch attacks?



: Wrong class detection

: Good detection

Possible scenarios of a classical adversarial attack on object detector.

Contribution

Definition of categories of evaluation criteria

Category	Setting	Description
Radiometric	Varying weather conditions Filters	Brightness, snow, rain, … JPEG transformations
Geometric	Rescaling Crop Affine transformations Distance w.r.t learning position	*** *** Rotations Shift from learning position
Transferability	Detector sensitivity Detector generalization	Sensitivity of a detector parameters to APAs Generalisation of an APA through multiple detectors

Table of evaluation settings by category and their brief description.

Example of settings

Example:

- Patch trained at top-left position

Measuring:

Category	Setting	Description
Radiometric	Varying weather conditions Filters	Brightness, snow, rain, … JPEG transformations

Category	Setting	Description
Geometric	Rescaling Crop Affine transformations Distance w.r.t learning position	*** *** Rotations Shift from learning position

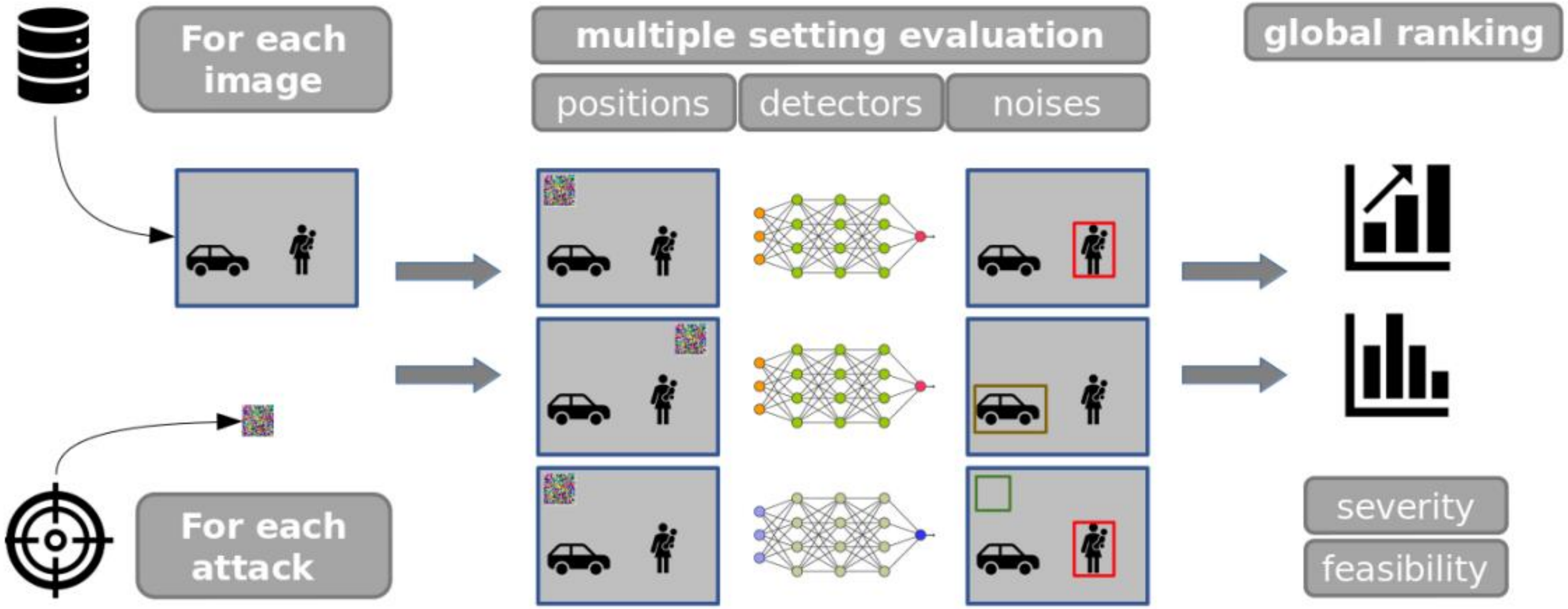
learning

test

test

Contribution (proposed evaluation pipeline)

- A framework to rank patch attacks



Structure of the proposed pipeline to evaluate APAs.

Experimental setup

Configurations:

- PASCAL VOC test dataset, YOLOv2 detector
- Evaluating patch contextual effects
- Attacking the *person* class
- Patch learned at top-left location and applied at the same position by default

Three state-of-the-art patch attacks:

- Dpatch (Liu *et al.*, 2018)
- Lee *et al.* (Lee *et al.*, 2019)
- Saha *et al.* (Saha *et al.*, 2020)

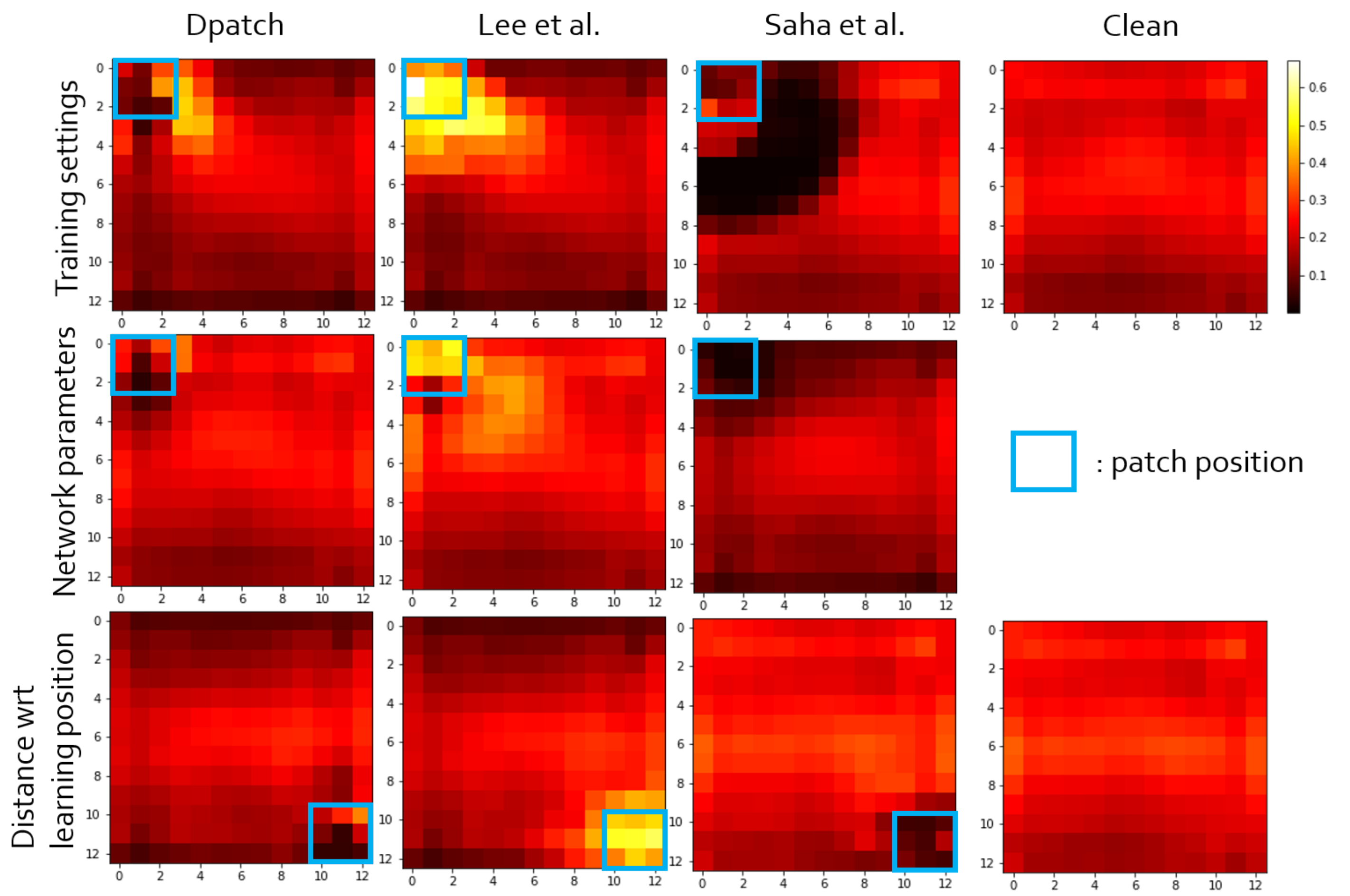
Experimental results (comparison table)

Setting	Attack	Attacked AP (%) with f.p	Attacked AP (%) without f.p	Cleaned AP (%)
Same as training	Dpatch	71.42	75.01	76.13
	Lee *et al.*	10.56	74.36	
	Saha *et al.*	**59.36**	**59.47**	
Other initialization	Dpatch	73.34	75.25	
	Lee *et al.*	60.35	75.42	
	Saha *et al.*	75.55	75.55	
Shift from learning position	Dpatch	70.61	77.87	80.01
	Lee *et al.*	53.02	78.73	
	Saha *et al.*	74.28	75.87	

Table of the evolution of the AP score for different setting evaluation and for different APA.

- **Better contextual effects for attack Saha et al. given training settings**
- **Dpatch and Lee *et al.* trying to be the salient object of images limiting their contextual effects**

Experimental results



Person class probability obtained by averaging anchors in cells over test set.

Two strategies:

- **Saha *et al.* : remove detections *i.e* reduce person class probability everywhere**
- **Dpatch and Lee *et al.* : create false alarms *i.e* increase person class probability around or on the patch**

Conclusion

- Our framework allows us to evaluate the real impact of APAs
- Comprehensive analysis of state-of-the-art adversarial patch attacks through a set of proposed evaluation settings
- Dpatch and Lee *et al.* have low contextual effects limiting their criticality
- Current attacks are sensitive to setting change, lowering the practical risk of current APA's

References

- [Song *et al.*, 2018] Dawn Song *et al.*, Physical adversarial examples for object detectors. In 12th USENIX workshop on offensive technologies (WOOT 18), 2018.
- [Saha *et al.*, 2020] Aniruddha Saha *et al.*, Role of spatial context in adversarial robustness for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 784–785, 2020.
- [Lee and Kolter, 2019] Mark Lee and Zico Kolter. On physical adversarial patches for object detection. Preprint arXiv:1906.11897, 2019.
- [Liu et *al.*, 2018] Xin Liu *et al.*, Dpatch: An adversarial patch attack on object detectors. SafeAI 2019 (AAAI Workshop on Artificial Intelligence Safety), 2018.

www.confiance.ai
contact@irt-systemx.fr