Academics hide humans from surveillance cameras with 2D prints

Senyang Jiang

Lab6

Automated surveillance

passive surveillance cameras

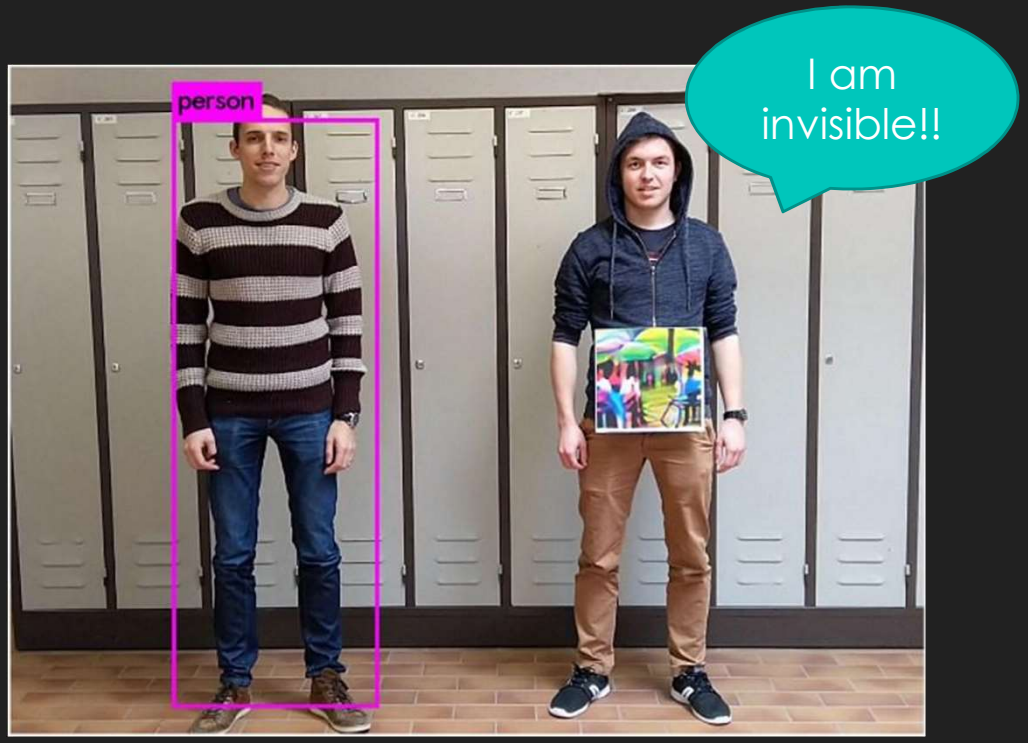
- ineffective
- labor and cost intensive
  - e.g. London Riot in 2011

Automated surveillance

- Uses machine learning algorithms to recognize humans in live video feeds
- Spot predefined 'suspicious' behaviors and send alarms
  - "AI guardman"

"AI guardman" Identifying shoplifter

Are they reliable enough?



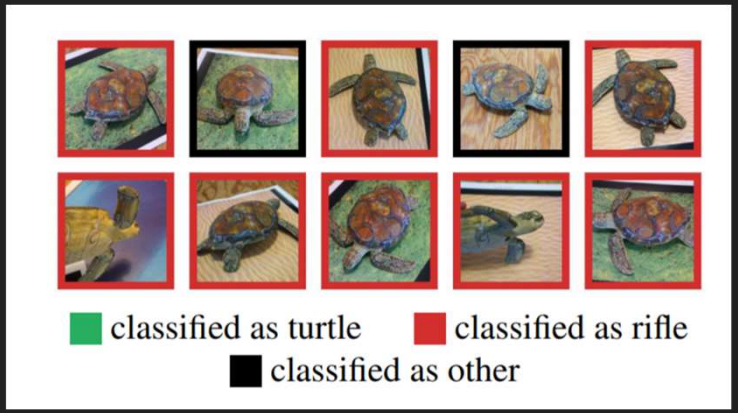
- Academics from a Belgian university have devised a method that uses a simple 2D image make wearers invisible to camera surveillance systems that rely on machine learning
- A form of **adversarial attack**
- The 2D image is called **adversarial "patch"**

Adversarial attack

- An adversarial attack basically means providing an ML model with a spurious input that will fool it into producing a wrong result.
- Recently, Deep neural networks have been found vulnerable to well-designed input samples
- The result can be some target class that the attacker wants to obtain.

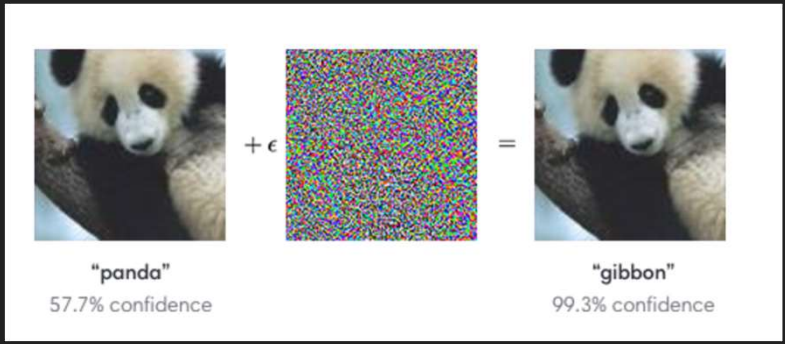
How?

- Two forms of Adversarial attack:
- Adversarial noise
- Adversarial patch



Adversarial noise

- Inputs: Image I, Target class C, Neural Network N
- Output: Modified Image I' that confuses N into predicting C
- Start with a random 'noise' filter X. Combine I & X and put it through N. Modify X so that two objectives are satisfied: 1. N predicts C for (I + X), and 2. X is as small in magnitude as possible.
- **Feature: the perturbations are designed to be subtle and hard to detect**

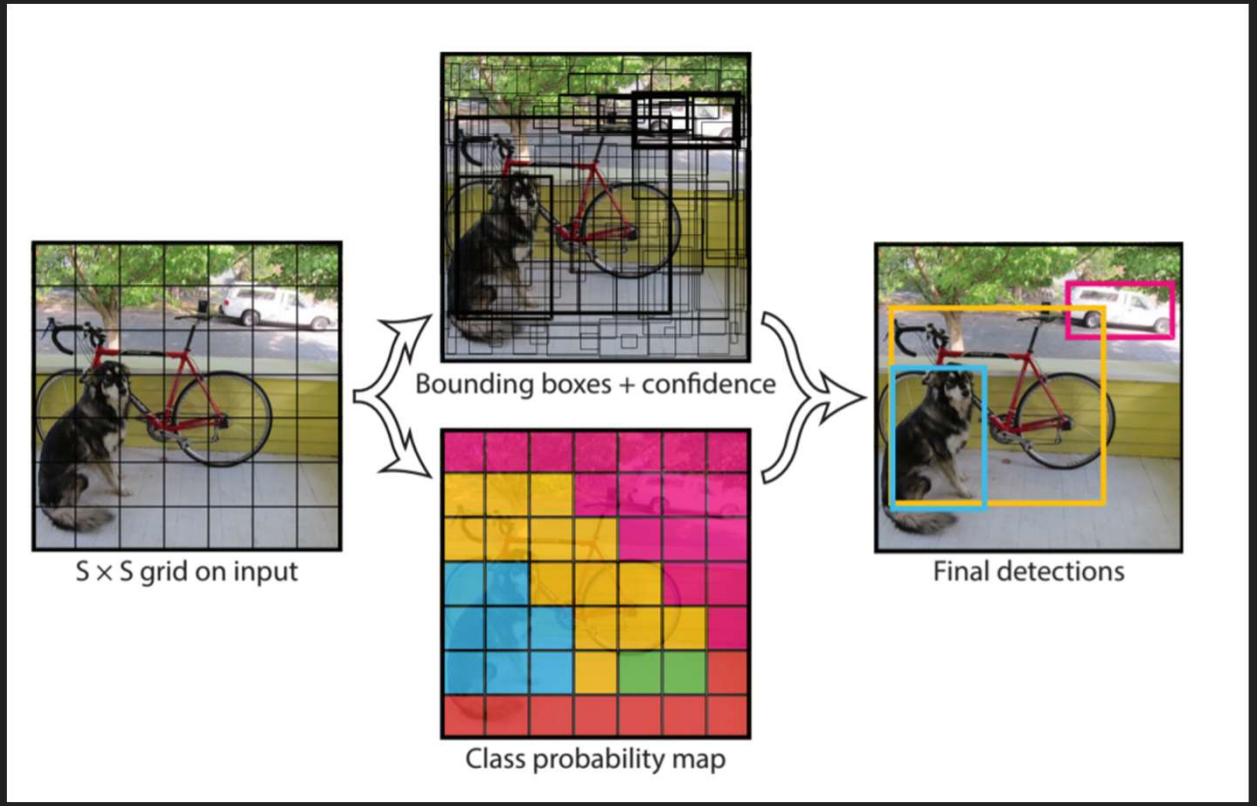


Adversarial patch

- Input: A target class C and Neural Network N (Note there is no input image here)
- Output: Patch P such that applying it to any image in any way, makes N predict C
- **Feature: this adversarial patch is "scene-independent"**
- Reason: a deep learning model will only detect the most "salient" item in an image

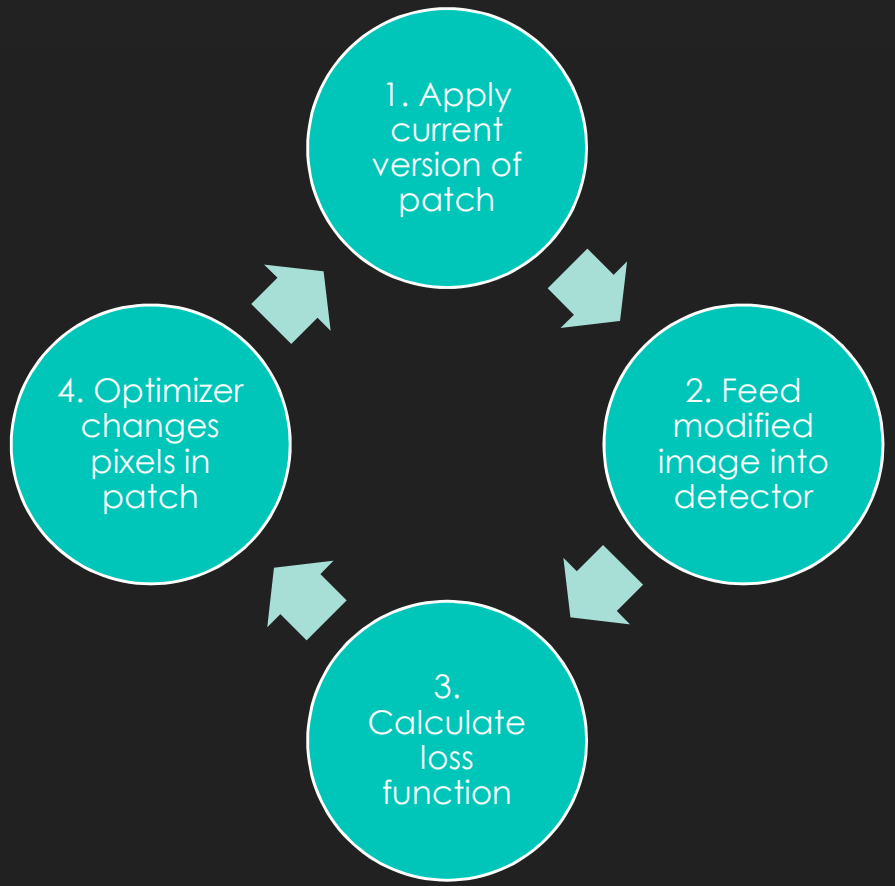
Attacks on surveillance cameras

- YOLOv2 object detector
- Uses Convolutional Neural Network
- Three predictions:
  - bounding boxes
  - object score
  - class scores

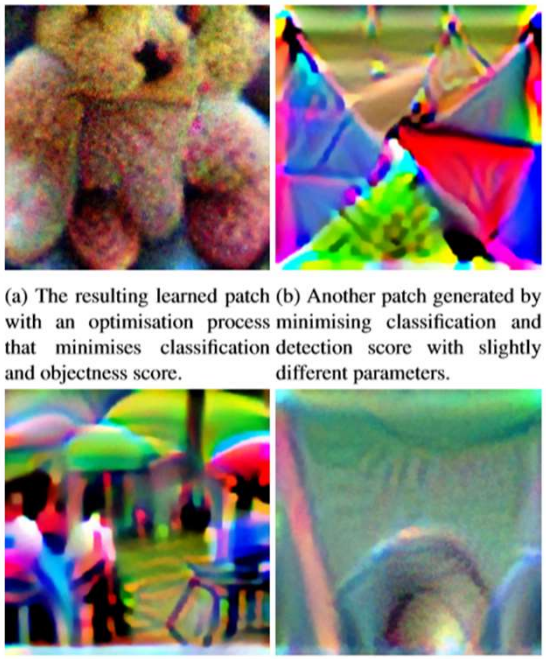


Attacks on surveillance cameras

- Three different approaches:
  - minimize the classification probability of class person
  - minimize the object score
  - A combination of minimizing both



Attacks on surveillance cameras



(a) The resulting learned patch with an optimisation process that minimises classification and objectness score.

(b) Another patch generated by minimising classification and detection score with slightly different parameters.

(c) Patch generated by minimising the objectness score.

(d) Minimising classification score only.

- Results: minimising the object score (OBJ) has the biggest impact on Average Precision

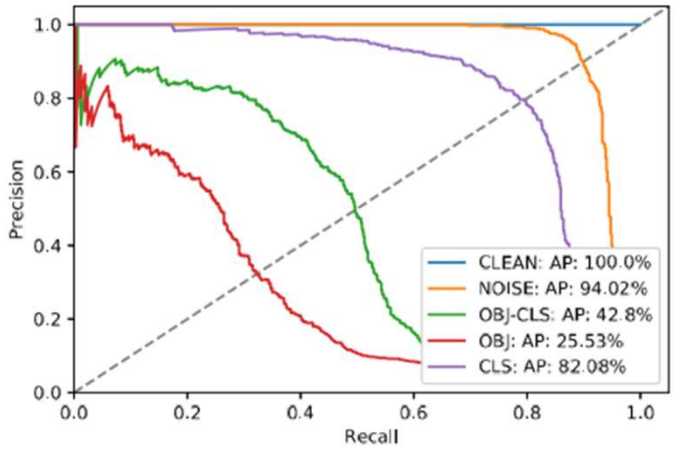
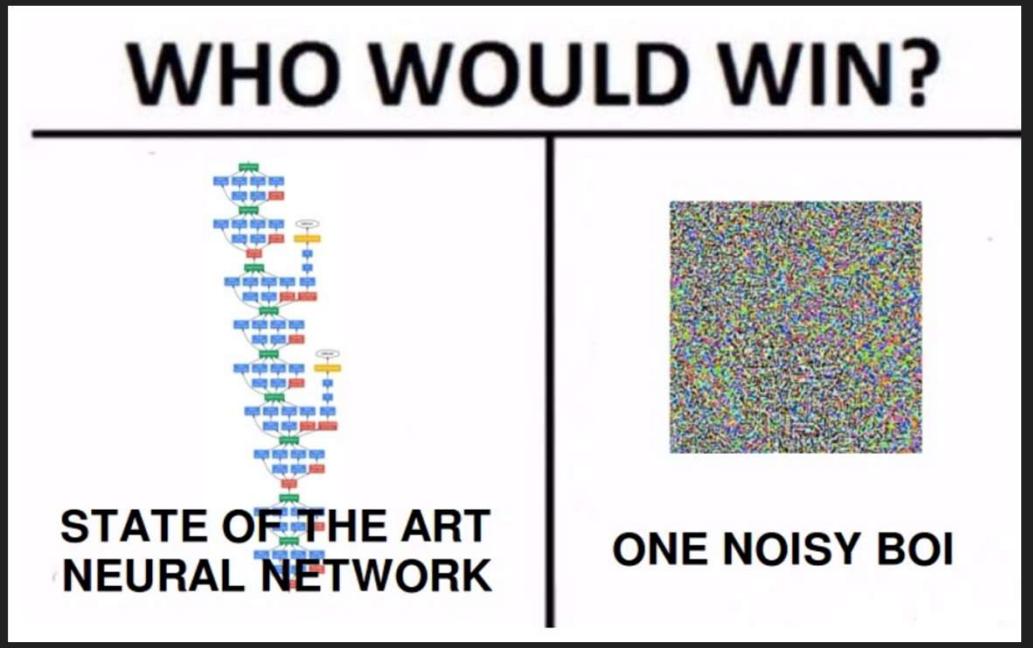


Figure 5: PR-curve of our different approaches (OBJ−CLS, OBJ and CLS), compared to a random patch (NOISE) and the original images (CLEAN).

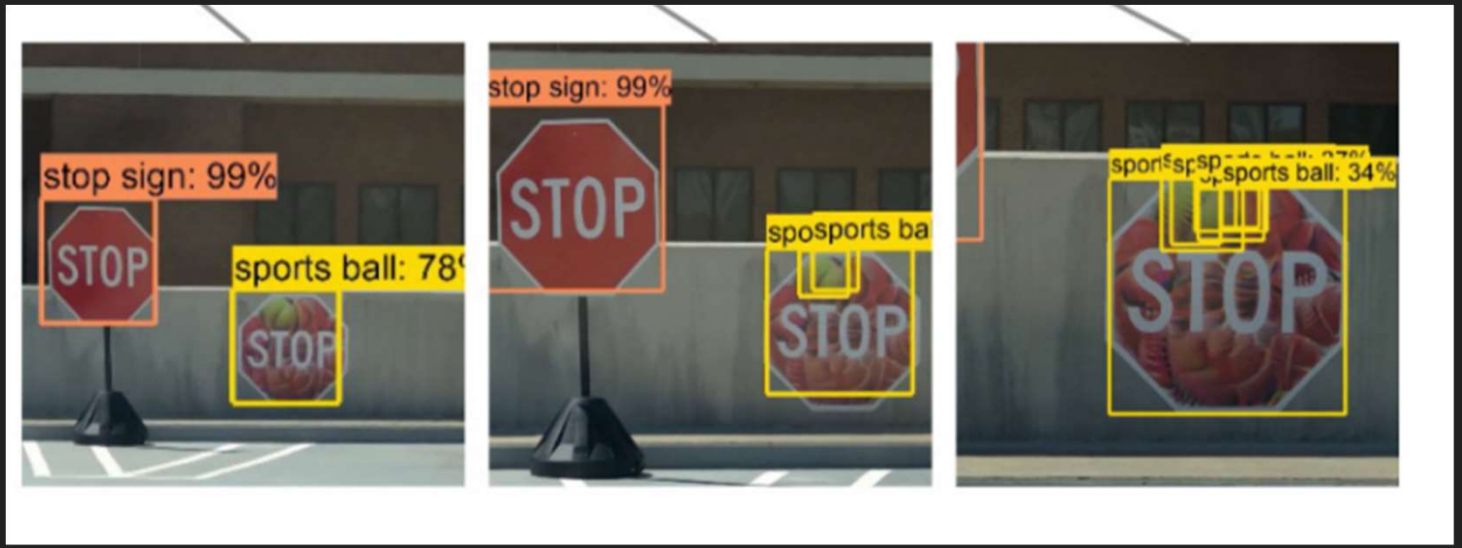
Evaluation

- A threat to the automated security systems
  - Intruders
  - Hiding certain objects



Evaluation

- Threats to other systems depending on Neural Network
  - e.g. Tesla Autopilot



Conclusion

"Things could get dangerous if thousands of self-driving cars rolling down the highway can only see toasters"

References

- J. Vincent, "This Japanese AI security camera shows the future of surveillance will be automated," The Verge, 26-Jun-2018. [Online]. Available: https://www.theverge.com/2018/6/26/17479068/ai-guardman-security-camera-shoplifter-japan-automated-surveillance . [Accessed: 20-May-2019].
- C. Cimpanu, "Academics hide humans from surveillance cameras with 2D prints," ZDNet, 05-May-2019. [Online]. Available: https://www.zdnet.com/article/academics-hide-humans-from-surveillance-cameras-with-2d-prints/ . [Accessed: 20-May-2019].
- S. Thys, W. V. Ranst, and T. Goedeme, "Fooling automated surveillance cameras: adversarial patches to attack person detection," arXiv:1904.08653v1, 2019.
- S. Joglekar, "Adversarial patches for CNNs explained," codeburst, 09-Jan-2018. [Online]. Available: https://codeburst.io/adversarial-patches-for-cnns-explained-d2838e58293 . [Accessed: 20-May-2019].
- S.-T. Chen, C. Cornelius, J. Martin, and D. H. Chau, "ShapeShifter: Robust Physical Adversarial Attack on Faster R-CNN Object Detector, " arXiv:1804.05810v3, 2019