Fooling thermal infrared pedestrian detectors in real world using small bulbs

Problem Formulation

- Problem formulation: $\min_{\delta} f_{obj}(\tilde{x}, \theta)$.
- Minimize output score of our detector by add perturbation in to original input image.
- To achieve our goal:
- - consider various image transformation and universal attack on different people.

$$\min_{\delta} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{t \in T} f_{obj}^{(i)}(\tilde{x}_t, \theta).$$

- Firstly, we calculate output score of detector on subject i when the input are images with various perturbation. Then we sum the scores of all N people and take average to show the generalization of attack on different pedestrains subjects.

Problem Formulation – Loss function

- L_{obj} represents max objectiveness score when the input is the patched image:

$$\min_{\delta} \frac{1}{N} \sum\nolimits_{i=1}^{N} \mathbb{E}_{t \in T} f_{obj}^{(i)}\left(\tilde{x}_t, \theta\right).$$

- L_{tv} represents total variation of the image, p is pixel value while (i , j) :

$$L_{tv} = \sum_{i,j} \sqrt{\left(p_{i,j} - p_{i+1,j}\right)^2 + \left(p_{i,j} - p_{i,j+1}\right)^2}.$$

- This loss function prevent the large pixel value between adjacent pixels. Therefore, the pixel value will change smoothly in different coordinate in image.

$$L = L_{obj} + \lambda L_{tv}.$$

For ensemble attack, in which case we want to minimize max objectiveness score L_{obj} of different detector at the same time. It means we need to minimize sum of ${L_{obj}}^{(i)}$.

$$L_{ensemble} = \sum_{i=1}^{M} L_{obj}^{(i)} + \lambda L_{tv}.$$

- Lastly, backpropagation is used to update patch iteratively.

Problem Formulation - Attack

Digital patch attack

1. Pixel level patch

- Goal: find the digital patch that could be easily realized using thermal material

2. The Gaussian function patch

- The Gaussian function fits temperature of a bulb at a horizontal line very well with a Root Mean Squared Error (RMSE) of 0.1511 only. The image patch composed of several bulbs in a cardboard is captured by infrared camera as a set of 2-D Gaussian function terms. Now the goal is that how can we arrange the image patch so that it could mislead pedestrian detectors.
- Since the amplitude and standard deviation of the Gaussian function is fixed to be measured values, the optimization parameter of each two-dimensional Gaussian function is the coordinate of the center point. It has much smaller number of parameters than pixel-level patch.

Problem Formulation - Attack

- \- Single Gaussian function:

Assuming that the pattern of a patch is superimposed by M spots that conform to Gaussian functions, where the center point of the i-th Gaussian function is (p_x, p_y), the amplitude amplification factor is s_i, and the standard deviation is σ_i. The measured s_i was 10.62, and σ_i was 70.07 in our experiment. We assume that the height of the entire image is h , the width is w, and the coordinate of a single-pixel is (x, y), where $x \in [0, w]$, $y \in [0, h]$, then the i-th Gaussian function is as follows:

$$g^{(i)}(x, y) = s_i \cdot \exp\left(-\frac{\left(x - p_x^{(i)}\right)^2 + \left(y - p_y^{(i)}\right)^2}{2\sigma_i^2}\right). \tag{6}$$

For the function of the image patch, it is the sum of background pixels and and all Gaussian function. We store the i th Gaussian function as a matrix whose dimensions is h * w, the value of each matrix element is just plug coordinate (x,y) in patch into the $g^{(i)}(x, y)$ function.

Physical board attack

Experiments and results

Dataset:

The paper selected 1011 infrared images containing people whose height is greater than 120 pixels from FLIR ADAS dataset . 710 of them as the training set and 301 as the test set => FLIR_person_select dataset

Target detector:

YOLOv3. Average precision = 0.9902 for training, 0.8522 for testing.

Simulation of physical attack:

- Pixel-level patch attack

YOLOv3 dropped by 74%, but the resulted patch contains lots of noise so that it is hard to realize.

- Gaussian functions patch attack

Patch is superimposed by multiple spots that conform to a two-dimensional Gaussian function. Various transformation like noise, rotation, translation, changes in brightness and contrast is designed to simulate real-life situation.

Experiments and results

- **Simulation of Physical Attack: .Gaussian functions patch attack**
- Input $\tilde{x}$: patched image with various transformation perturbation

X is original image, after we add P_{syn}, the total perturbation/patch, it become patched image $\tilde{x}$

$$P_{syn} = P_{back} + \sum_{i=1}^{M} G_i$$

$$G_i = \begin{pmatrix} g^{(i)}(0,0) & \cdots & g^{(i)}(0,w) \\ \vdots & \ddots & \vdots \\ g^{(i)}(h,0) & \cdots & g^{(i)}(h,w) \end{pmatrix}.$$

- Loss function:

$$\min_{\delta} \frac{1}{N} \sum\nolimits_{i=1}^{N} \mathbb{E}_{t \in T} f_{obj}^{(i)}(\tilde{x}_t, \theta). \qquad L_{tv} = \sum_{i,j} \sqrt{(p_{i,j} - p_{i+1,j})^2 + (p_{i,j} - p_{i,j+1})^2}. \qquad L = L_{obj} + \lambda L_{tv}.$$

The loss function updated the patch pixel matrix, not detector; detector provide f, theta, which won't be updated.

- Optimization: Backpropagation to minimize the loss function L, so that we can find
- Optimizer:stochastic gradient descent
- Detector: pre-trained YOLOv3

=> Lastly, optimized patch is obtained.

Experiment and results

To realise the attack performance of obtained optimized image, the experiment is conducted as follow.

Control experiment for comparison:

Random noise patches with maximum amplitude value 1 and constant pixel value patches (blank patches)

Experiments & Result:

- Performance compared to control image: the Gaussian functions patch we designed made the average precision (AP, the area under the PR curve) of the target detector drop by 64.12, compared to 25.05% and 29.69% for random noise patch and blank patch, respectively.
- Performance of patched image with different spots(Gaussian function terms):

22 spots have good attack effect

The number	The AP dropped by
9	46.02%
15	51.26%
22	64.12%
25	65.74%
36	66.88%

Experiment and results

- The effects of patch size

The larger the image is, the better the attack effect => limitation of the patch attack method.

- Evaluation of the attack in real life

1. Economic feasibility: cost less than 5 dollar.
2. Performance of detector when pedestrians Holding designed physical board/blank board/no board:

The result showed that the cardboard caused the average precision (AP) of the target detector to drop by 34.48%, while a blank board with the same size caused the AP to drop by 14.91% only.

- Ensemble attack : Transferability evaluation

A new Gaussian patch is trained by combination of YOLOv3, Faster-RCNN, and Mask-RCNN detectors to improve transferability. Compared to single detector, attack effect and transferability improve a lot.

Train \ Test	Cascade RCNN	RetinaNet
YOLOv3	11.60%	25.86%
YOLOv3+Faster -RCNN+Mask-RCNN	35.28%	46.95%

Conclusion

Objective & Works:

The paper proposes a physical attack method with small bulbs on a board against the state of-the-art pedestrian detectors. Our goal is to make infrared pedestrian detectors unable to detect real-world pedestrians.

Experiments & Results:

YOLOv3. The average precision (AP) dropped by 64.12%， blank board with the same size caused the AP to drop by 29.69% only. Real world experiment: In recorded videos, the physical board caused AP of the target detector to drop by 34.48%, while a blank board with the same size caused the AP to drop by 14.91% only.