Exploring the Vulnerability of Single Shot Module in Object Detectors via Imperceptible Background Patches

Yuezun Li 1 Xiao Bian 2 Ming-Ching Chang 1 Siwei Lyu 1

1 University at Albany, State University of New York, Albany, New York, USA
2 GE Global Research Center, Niskayuna, New York, USA

Abstract

We explore vulnerability of the Single Shot Module (SSM) commonly used in recent object detectors, by adding small perturbations to patches in the background outside object of interest to attack the object detection task.

Backgrounds

• Single Shot Module (SSM) refers to (1) the Region Proposal Network (RPN) commonly used in two-stage object detectors, or (2) the single-stage object detector itself.
• Adversarial perturbations are intentionally designed noises that are imperceptible to human observers yet can seriously harm the performance of a deep neural network.

Overview

SSM Object Detector Attack Pipeline

(a) Input image. 
(b) Background patches generated from our method. 
(c) SSM base-network (RPN of two-stage detectors or the single-stage detector).
(d) Output of SSM. Our attack can effectively disrupt the top ranked results by decreasing true positives (black) and increasing false positives (red).
(e) Top ranked object proposals (for two-stage detectors) or detection results (for single-stage detectors).
(f) Sub-network of two-stage object detectors for class labels prediction and shape refinements.

Methods

Problem Formulation

The proposed SSM adversarial attack is to search for suitable background patches in terms of geometry (location, size and shape) and pixel changes to be altered.

Given input image $I$, we formulate this optimization as the minimization of three loss terms:

1. True Positive Class (TPC) loss: $L_{tpc}$
2. True Positive Shape (TPS) loss: $L_{shape}$
3. False Positive Class (FPC) loss: $L_{fpc}$

$$\min_{I \in Q} \left[ L_{tpc}(I \cup Q, F) + L_{shape}(I \cup Q, F) + L_{fpc}(I \cup Q, F) \right],$$

s.t. $\text{PSNR}(I \cup Q) \geq \varepsilon$. Location and shape of the background patches $Q$ and the included pixel value in image $I$.

Background Patches Generation

Algorithm 1 Background Patch Generation

Require: SSM model $F$; input image $I$; maximal iteration $T$
1: $Q_0 = I; \ v = 0$
2: while $I < T$ and $\sum_{i=1}^v |Q_i| \neq 0$ do
3: $Q_i = Q_{v} \cap \min_{Q \in \text{PSNR}}(I \cup Q, F)$
4: if $v = 0$ then
5: $Q_0 = \text{initial background patches}$
6: else
7: $Q_i = \text{expanded background patches}$
8: $Q_{v+1} = \text{expanded background patches}$
9: $P_i = \text{PSNR}(Q_i)$
10: $Q = \text{expanding direction where gradient intensity increases the most}$
11: if $\text{PSNR}(Q_i) < \varepsilon$ then
12: break
13: else
14: $r = r + 1$
15: Ensure: Adversarial perturbed image $I_v$

Results

Experiment on 5 two-stage object detectors with 5 different Region Proposal Networks (RPNs) and 8 single-stage object detectors at mAP at $t = 0.5 / 0.7$ (1st/2nd value in table).


Conclusion

- The proposed SSM background-patch attack can effectively harm mainstream deep object detection networks by only altering imperceptible pixels in the background that results in significantly decreased true positives and increased false positives.
- Experiments on mainstream object detectors expose such vulnerability.