SLAP: Improving Physical Adversarial Examples with Short-Lived Adversarial Perturbations

Giulio Lovisotto† (presenting)
Henry Turner†, Ivo Sluganovic†, Martin Strohmeier*, Ivan Martinovic†

†University of Oxford, *armasuisse

USENIX Security ’21
Motivation

• Multiple deep neural networks
• Multi-tasks
• Multiple predictions
• Safety & Security?

How can we alter a scene so that the camera-perceived inputs deceive a neural network?
Threat Model – Attack Vector

- One attack vector: Adversarial patches
  - Detectable behavior (see SentiNet$^3$)
  - Static scene modification (conspicuous)

- Any other way?
  - What if we used an RGB projector

- Perturbation

References:

[1] - Eykholt, K. et al., 2018. Robust physical-world attacks on deep learning visual classification. CVPR.
[2] - Zhao, Y. et al., 2019. Seeing isn’t believing: Towards more robust adversarial attack against real world object detectors. ACM CCS.
Method – Color Realizability

• Learn color space with a neural network
  • overcomes limit of Non-Printability Score\(^4\) used for patches

• Collect (i) surface color, (ii) projected color, (iii) resulting color with camera

\[
\text{Loss}_P = \arg \min_{\theta_1} \sum_{\forall c_s, c_p} \left\| \mathcal{P}(c_s, c_p) - c_o^{(s,p)} \right\|_1
\]

Method – AE Generation

- Real-world AE have to be robust to environmental changes
- Use *Expectation-over-transformation* with data augmentation
Real-world Evaluation

Evaluation on 4 target networks: LISACNN, GTSRBCNN, YOLOv3, MaskRCNN

- Controlled lighting indoor
- Outdoor in moving vehicle settings

Outdoor - 5.30pm - ~200lx
Results – Attack Success

Mis-detection rate (no. of times a stop sign is not detected)

- **YOLOv3 (120lx)**
  - Baseline mis-detection
  - SLAP
Results – Attack Success

Increasing ambient light

600 lx

300 lx

120 lx

YOLOv3  MaskRCNN  GTSRBCNN  LISACNN
Results - Defenses

- Adversarial learning
  \[ \hat{J}(\theta, x, y) = \alpha J(\theta, x, y) + (1 - \alpha) J(\theta, x + \epsilon \text{sign} (\nabla_x J(\theta, x, y))) \]

- Input randomization

- SentiNet (specifically designed for physical AE)

---

<table>
<thead>
<tr>
<th>Network</th>
<th>Ambient Light (lx)</th>
<th>Attack Success</th>
<th>Adversarial Learning</th>
<th>Input Randomization</th>
<th>SentiNet Random</th>
<th>SentiNet Checkerboard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gtsrb-CNN</td>
<td>120</td>
<td>99.96%</td>
<td>20.23% (79.73%)</td>
<td>99.55% (0.40%)</td>
<td>93.43%</td>
<td>95.45%</td>
</tr>
<tr>
<td></td>
<td>180</td>
<td>90.53%</td>
<td>23.57% (66.97%)</td>
<td>90.02% (0.51%)</td>
<td>93.19%</td>
<td>93.72%</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>56.51%</td>
<td>48.18% (+30.27%)</td>
<td>86.78% (-16.33%)</td>
<td>96.97%</td>
<td>96.46%</td>
</tr>
<tr>
<td></td>
<td>440</td>
<td>56.34%</td>
<td>40.24% (+16.10%)</td>
<td>82.96% (+26.61%)</td>
<td>95.81%</td>
<td>96.34%</td>
</tr>
<tr>
<td></td>
<td>600</td>
<td>12.79%</td>
<td>10.91% (-1.88%)</td>
<td>51.37% (+38.58%)</td>
<td>95.29%</td>
<td>95.29%</td>
</tr>
<tr>
<td>Lisa-CNN</td>
<td>120</td>
<td>100.00%</td>
<td>0.06% (99.94%)</td>
<td>100.00% (+0.00%)</td>
<td>94.24%</td>
<td>95.29%</td>
</tr>
<tr>
<td></td>
<td>180</td>
<td>99.95%</td>
<td>0.88% (99.07%)</td>
<td>99.90% (-0.05%)</td>
<td>100.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>99.81%</td>
<td>0.00% (99.81%)</td>
<td>99.98% (+0.17%)</td>
<td>94.76%</td>
<td>96.86%</td>
</tr>
<tr>
<td></td>
<td>440</td>
<td>98.44%</td>
<td>0.59% (97.85%)</td>
<td>99.95% (+1.51%)</td>
<td>100.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td></td>
<td>600</td>
<td>69.05%</td>
<td>0.04% (69.01%)</td>
<td>95.71% (+26.67%)</td>
<td>95.81%</td>
<td>96.86%</td>
</tr>
</tbody>
</table>

Results - Generalizability

Extension to other models and objects:

- YOLOv3
- MaskRCNN
- GTSRBCNN
- LISACNN

 imágenes of objects detected by each model with corresponding confidence scores:
- Toothbrush (0.13)
- Bottle (0.67)
- Bottle (0.66)
- Give Way (1.00)
- Give Way (0.98)
- Speed Limit 45 (0.54)
- Work in Progress (0.91)
Conclusion

• Demonstrated that SLAP works robustly in real-world for both image classification (traffic signs) and object detection

• No straightforward defense detects SLAP
  • Adversarial learning might end up in arms race

• Introduce new capabilities for AE-based attackers: alter scene content dynamically
  • Highlights the practical need of new type of adversarial robustness for sequential data
Thanks

Thanks for your attention!

Contact: giulio.lovisotto@cs.ox.ac.uk

SLAP project page https://github.com/ssloxford/short-lived-adversarial-perturbations/