CS 6220 - Big Data Systems and Analytics

Saurab Sirpurkar
Assignment 2

24 September 2021

A critique on
Adversarial Objectness Gradient Attacks in Real-time Object Detection Systems

TOG is an assemblage of adversarial attacks on Deep Neural Networks that includes untargeted random attacks, and three specific targeted attacks namely object-vanishing, object-fabrication and object-mislabeling, making it an extremely versatile attack generation system. It lets attackers configure adversarial parameters to mislead the model towards *specific* misclassifications. Therefore, it can be used to launch sophisticated attacks on self-driving cars, face-recognition systems, medical diagnostic systems etc.

Many previous papers focused on adversarial examples for two-phase models. But TOG can also target one-phase detection algorithms. Also, its algorithm can train offline on a training set to later make these attacks with negligible perturbation delay on the fly. TOG further exposes vulnerabilities in these deep learning models by generating the same attacks by perturbing only a small patch in the image (TOG-patch). For example, a car can attach one of these perturbed stickers that TOG generates and confuse the Tesla driving behind into thinking its a wall. This comes under object-misclassification. Similarly, a sticker on the stop sign could make it invisible to the Tesla. This comes under object-vanishing. Creating certain markings on the sensor can lead it to detecting ghost pedestrians thereby having to brake abruptly. This comes under object-fabrication. Clearly, we must tackle these issues, and for that, we must study how TOG operates.

Deep Learning models update their parameters progressively to minimise the difference between the estimated labels and the correct ones. Adversarial models work backwards. They progressively manipulate the image until the predictive model gets confused between two labels enough to start misclassifying. The direction of this perturbation can be obtained by calculating the sign of the gradient of the adversarial loss function. Iteratively, with slight and often indiscernible perturbations, benign images turn adversarial. The following image shows an example of mislabeling perturbation generated by TOG which follows a similar algorithm.

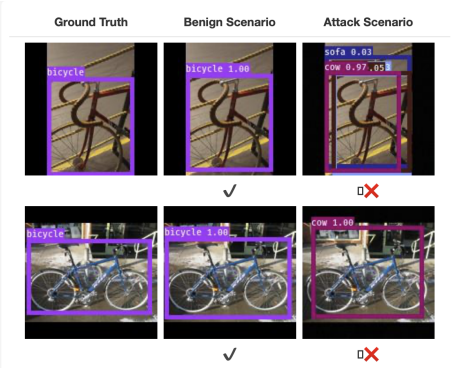

Figure 1: TOG perturbation leading to mislabeling of an object
Source: security4ai-vlab applet, Distributed Data Intensive Systems Lab, Georgia Tech.

TOG reports high rates of success across multiple evaluation metrics. For example, a vanishing TOG patch can confuse object detectors by vanishing them 97% of the time against a YOLOv3-D detector. It outperforms all other attack systems like UEA and RAP on FRCNNs for untargeted attacks, and it also extremely successful in all three of its targeted attacks. It has a high transferability to SSD models compared to other attacks.

This paper on TOG showed us three kinds of targeted attacks and their evaluation in depth. However, the authors could have cemented TOG's benefits by providing a methodology to do all three targeted attacks *simultaneously* and evaluating it. Also, the TOG patches expose serious model vulnerabilities by targeting only portions of images, but they are not indiscernible to the human eye. Therefore, for digital patches, instead of the proposed randomized initialization, initialising the pixels identical to the benign image might lead to more curious discussion on what is the smallest (indiscernible) perturbation area that TOG attacks need to succeed. Moreover, discussion on TOG's efficacy when facing popular adversarial defense mechanisms and suggestions on defense-specific adaptations could further prove why it is an extremely strong adversarial attack generation framework.