Adversarial examples in Autonomous Mobile Robots: A Survey

Prabhant Singh
University of Tartu
Estonia
Prabhant.singh@ut.ee

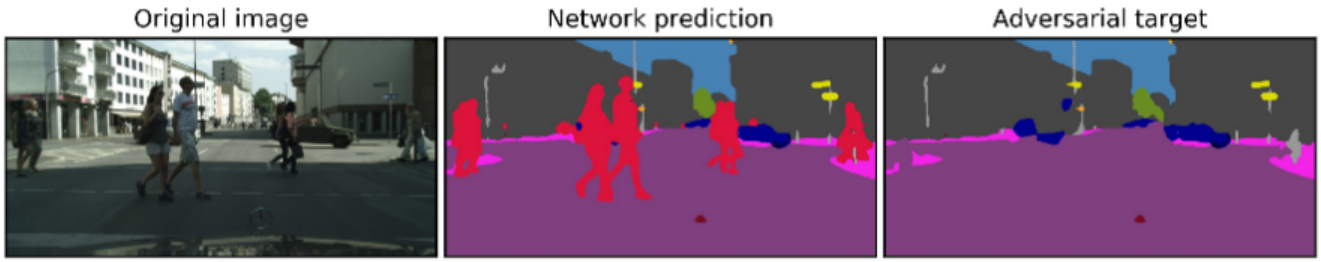

Figure 1. illustration of an adversary generating a dynamic target segmentation for hiding pedestrians[11]

Abstract

Autonomous mobile robots (AMR, e.g. autonomous cars and drones) have gathered much attention lately. Adversarial examples in machine learning [6], i.e. carefully perturbed input samples designed to fool a machine learning system, have raised wide concerns about the security and safety implications of relying on machine learning decisions when deployed in the wild.This paper provides an overview of the evolution of adversarial examples in the area of autonomous mobile robots. This work also introduces a point based metric system on the robustness and limitations of adversarial examples in real world.

Keywords Adversarial examples, autonomous mobile robots

1 Introduction

Most of the Vision based systems use Neural network based systems for classification detection and segmentation. In most of the cases Convolutional neural networks are used for classification and detection tasks(using VGG and Imagenet architecture), Neural network architectures with FCNs are used for the segmentation based tasks.

I will be using the terminology described by Battista et al[6] here. Most of these tasks utilize neural networks for the state of the art performance hence I will be limiting our work on adversarial examples developed for neural networks not classical machine learning algorithms.
The tasks for which autonomous vehicles use learning algorithms are:

1.1 Image classification

The AMVs use Convolutional neural networks to classify the traffic signs, objects, traffic lights and the road type. Convolutional neural networks demonstrated state of the art performance in the image classification problem. The networks usually use a VGG/Imagenet like architecture.

1.2 Object detection:

Object detection and identification is a critical factor in the working of autonomous vehicles. The object detection task

uses deep neural networks to achieve reasonable results which ensures that the object detection can be used in real-world scenarios.

1.3 Depth estimation:
Many recent works like [20]perform depth estimation using the convolutional network. In [20] the authors introduce Monodepth, an encoder-decoder architecture which regresses the depth of given monocular input image.

1.4 Semantic segmentation:
Semantic segmentation is a critical part of the operation of autonomous vehicles. Semantic segmentation allows the AMV to segment the various elements in the view like roads, pedestrians, sidewalks, man-made objects, trees, etc. state of art semantic segmentation techniques can be developed by using deep neural architectures with FCNs.

2 Background on Adversarial examples

Adversarial examples were first developed in 2004 by Dalvi et al[8] and immediately later by Lowes and Meek who studied the problem of adversarial setting in the context of spam filtering. the first adversarial examples in the context of neural networks in 2014 by Szegedy et al[8] where a carefully crafted noise was introduced in the image and then inserted on inception model trained on imagenet.
The initial work showed that adversarial examples only exists in laboratory settings though recent work has demonstrated that adversarial examples exist in physical realm too.

2.1 Ways to attack a system with adversarial examples
Adversarial examples can be used to attack a system in 2 ways[6]

- **Train time attack or poisoning attack:** Poisoning attacks refer to injecting adversarial examples in a model while training so that the model causes misclassification of several elements after deployment. This kind of adversarial examples can be quite hard to detect.
- **Test time attacks/evasion attacks:** Test time attacks refer to the attacks which were carried out on a deployed model. These are the most common attacks in the field of adversarial machine learning as these can be deployed and tested with any model without any specific permissions.

2.2 Goals of crafting adversarial examples:

The 3 major goals of crafting adversarial examples are

- **Targeted misclassification:** The targeted misclassification refers to the scenario where the attacker can supply input to the model and get the incorrect output as the desired class as specified by the attacker. In terms of autonomous vehicles, it can mean generating an adversarial example which can cause the removal of a specific entity from the landscape like removing pedestrians.
- **Untargeted misclassification:** Untargeted misclassification refers to the scenario where the attacker can cause misclassification of input to any other class except the true class. For example, causing stop signage to be classified as any other or segmenting the cars as some other.
- **Confidence reduction:** As the name states confidence reduction can be caused by adversarial examples, this is the most common consequence of generating adversarial examples in the black box setting.

2.3 Adversarial examples in terms of attacker knowledge

I can classify the domain of adversarial attacks in terms of attackers knowledge :

White box: In white box attacks, the attacker is assumed to know everything about the targeted systems ie training data feature set learning algorithm and objective function. This setting allows an attacker to craft worst case adversarial examples that are targeted misclassification. Though white box attack might sound unrealistic but is quite possible assuming the fact that most of the applications use open datasets for training like imagenet for classification tasks or Pascal VOC, kitti, cityscapes for semantic segmentation tasks. Most of the learning algorithm is also known as applications use pre-trained network weights to save computation expense and time.

Grey box: Grey box or limited knowledge attacks can be assumed to know the few details of the model like learning algorithm used or dataset used but might not know the hyperparameter values.

Blackbox: Black box attacks assumes that attacker knows nothing very little about the system. Recent work has shown

that machine learning applications can be threatened by any substantial knowledge of the feature space, learning algorithms or training data. This can be done be wither querying the system in a black box manner to get feedback and confidence scores. Another approach to constructing adversarial examples in a black box setting is to craft transferable adversarial examples which can mean that an adversarial example was crafted for a model but is also threatening to another model in a black box setting. This is a major concern as there are very few ways to prevent against the adversarial examples in black box setting. In some of the cases the attacker knows a bit about the systems so the distinction between blackbox and greybox in that scenario is still a bit ambiguous.

3 Adversarial examples in Autonomous Mobile Robots

3.1 Metrics

This work aims to introduce a Impact based category system for the robustness of the adversarial examples in autonomous vehicle scenarios:

Impact	Explanation
Low	Adversarial examples are not deployable in real world and are not robust under the black box setting and have low impact in the real world scenario
Medium	Adversarial examples are robust under physical noise without major perturbations. or Adversarial examples are valid under blackbox scenarios in controlled physical setting and have high impact in real world scenario.
High	Adversarial examples are transferable and robust under most of the physical noise without major visible perturbations and have serious consequences in real world

This section covers the adversarial examples developed in various tasks which uses learning based algorithms.

3.2 Image classification:

The use of adversarial examples in image classification can be segregated in two subtopics:

- Adversarial examples for the entire image in the scenario as used by kitticlass
- Adversarial examples on traffic sign detection and classification.

3.2.1 Road sign classification problem

Road sign classification problem is one of the most common benchmark to develop adversarial examples in image classification task as they play an important role in transportation safety, as well as they can be modified by an attacker who might not have control over vehicle's systems.

The robustness of adversarial examples in image classification task of autonomous vehicles is quite debatable[4]. The main challenge while crafting adversarial examples in the physical world are:

1. Physical cameras might not capture the minute and sensitive perturbations.
2. Varying backgrounds in training and testing of perturbations.
3. Errors in fabrication process of the perturbed examples.(errors in printing)
4. Angles and speed of the the camera on autonomous vehicles might not be able to capture the effect of the perturbations.

3.2.2 Adversarial examples for static targets:

There has been a lot of research in generating physical adversarial examples for traffic signs with the static positioning of the camera[3,10]. The generation algorithm usually consist of a optimization problem to maximize the loss function. . The threat of adversarial examples in static positioning of the camera is still low as most of the adversarial examples loses these robustness during the motion of the camera but it serves as a good benchmark for the further research.

3.2.3 Adversarial examples for dynamic targets:

The recent work on generating adversarial examples for the dynamic positioning of the camera[3] has been quite successful. This work disproved the previous assumptions[4] where adversarial examples might lose their robustness due to the movement of the camera. The Robust Perturbations attack[3] has been the most robust attack among the work done yet with the accuracy of 80-100%(accuracy for different approaches).

These kind of adversarial examples can be a threat to autonomous vehicles in real world as shown by the recent

work[3,10]. For examples a stop sign can be misclassified as Speed45mph, which can be a threat to the operation of the adversarial examples. This kind of adversarial examples crafted with Robust Perturbation algorithm are even robust under various real world noise like camera rotation, moving vehicle, cropping etc. These adversarial examples are robust under most of the physical noise:

$Y_{true} = Yt_{target}$ for every angle θ at any speed ν

Threat level: 8 - The current state-of-the-art adversarial examples were able to fool the machine learning systems in real-world setting under any physical noise, the adversarial examples developed were transferable and can be deployed by an attacker which can cause a severe harm to the performance of the Autonomous vehicle. The research still lacks testing of robustness

3.3 Object detection:

The network for object detection in the autonomous vehicles can be attacked in following ways

- Object not detected
- More objects than the desired objects are detected

In [1,5] the authors showed that adversarial examples are a threat to the network in object detection where the object is misclassified. One important aspect of this algorithm[1,5] is that it generates adversarial examples even in the black box scenarios which can be highly risky for the autonomous vehicles.

Though Dense adversary generation algorithm[1] is transferable under various networks, datasets and even supports cross task transfer(adversarial examples crafted for object detection is also robust in object detection and segmentation domain) but it is only robust under digital domain. Whereas the adversarial examples crafted by [5] were crafted exclusively for physical domain.

The robust adversarial examples under real world scenarios were developed by [5] which completely fooled the classifier under certain possible conditions. I would not suggest that adversarial examples are 100% robust under physical domain as the approach[5] crafts large visible perturbations.

Threat level : 6 - Adversarial examples were successfully crafted in the area of object detection which can hinder the performance of the AMRs in real world scenario. The main limitation of these adversarial examples were that the perturbations were highly visible and the image almost becomes unrecognisable in cases of black box scenarios.

3.4 Semantic segmentation:

Adversarial examples in the domain of semantic segmentation task of the autonomous vehicles are produced in very recent work.

For semantic segmentation, the possible scenarios for generating adversarial examples are:

1. Static target segmentation: Motion of the camera is static.
2. Dynamic target segmentation: The camera is dynamically positioned.

There has been no work on developing adversarial examples for semantic segmentation in real world, nor there has been any suggested approach. The current threat from these adversarial examples only exists in Digital domain. Most of the work in developing adversarial examples in semantic segmentation task uses Universal adversarial perturbations denoted as Ξ [9,16,11].

Adversarial examples for Static target segmentation
In this scenario the adversary defines a fixed segmentation such that:

$y_{true} = y_{pred} + \Xi$ for a certain time frame t

to hide any suspicious activity for a certain time span. These examples are a potential threat to the digital systems, for example an attacker omitted a pedestrian[11] from the road which can cause a threat.

Adversarial examples for Dynamic target segmentation

The adversarial examples for dynamic targets were first demonstrated by [11] using universal adversarial perturbations. In case of motion the static adversarial examples would fail as they were not crafted for changing scenes. In contrast the adversarial examples for dynamic targets focus on keeping the network segmentation unchanged so that

$Y_{true} = Y_{pred} + \Xi$ for all values of (i,j) .

The real world adversarial examples for semantic segmentation still remains a challenge in real world scenarios as there is no possible way currently to deploy

these adversarial examples on real world scenarios without the attacker's access on the system.

Threat level: 4 - Adversarial examples only exists in the digital realm and there is no possible way to deploy these examples in the real world yet. Hence the threat of adversarial examples in semantic segmentation is still low.

3.5 Depth estimation:

The adversarial examples for depth estimation were proposed by [19] using universal adversarial perturbations were able to develop scenarios which were also transferable on different datasets and different networks such that:

$D_{target} + \Xi = D_{true}$

but like semantic segmentation it still remains an open question to deploy these adversarial examples in real world scenarios.

Though there has not been any research in dynamic adversarial examples for depth estimation like [11] in case of semantic adversarial examples.

Threat level: 2 - Adversarial examples only exists in the digital domain and only influenced the static images. Hence the adversarial examples in the domain of depth estimation can be considered to have a low impact.

4. Conclusion

In this work I have presented a brief overview of work related to security of the autonomous mobile robots, with the goal of summarizing the challenges in the deployment of autonomous robots using learning based algorithms. An impact based system has been introduced to determine the robustness of the adversarial examples in autonomous mobile robots.
This paper concludes that adversarial examples in the field of image classification and object detection are a real world threat to the deployment of these systems. Adversarial examples in semantic segmentation and depth estimation tasks are not deployable in real-world yet.I conclude this paper by suggesting to use a common evaluation metric to measure the robustness as mentioned in this paper.

References

[1] Xie, Cihang and Wang, Jianyu and Zhang, Zhishuai and Zhou, Yuyin and Xie, Lingxi and Yuille, Alan, Adversarial Examples for Semantic Segmentation and Object Detection,International

[2] Anish Athalye, Logan Engstrom, Andrew Ilyas, Kevin Kwok , Synthesizing Robust Adversarial Examples. arXiv:1707.07397

[3] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, Dawn Song, Robust Physical-World Attacks on Deep Learning Models. CVPR 2018.

[4] Jiajun Lu, Hussein Sibai, Evan Fabry, David Forsyth, NO Need to Worry about Adversarial Examples in Object Detection in Autonomous Vehicles, CVPR 2017.

[5] Jiajun Lu, Hussein Sibai, Evan Fabry, Adversarial Examples that Fool Detectors, arXiv:1712.02494 .

[6] Alexey Kurakin, Ian Goodfellow, Samy Bengio, Adversarial examples in the physical world, arXiv:1607.02533.

[7] Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy, Explaining and Harnessing Adversarial Examples, ICLR 2015.

[8] Battista Biggio, Fabio Roli, Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning, arXiv:1712.03141.

[9] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard, Universal adversarial perturbations, CVPR 2017.

[10] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, Prateek Mittal, DARTS: Deceiving Autonomous Cars with Toxic Signs, arXiv:1802.06430.

[11] Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, Volker Fischer, Universal Adversarial Perturbations Against Semantic Image Segmentation, arXiv:1704.05712.

[12] Moustapha Cisse, Yossi Adi, Natalia Neverova, Joseph Keshet, Houdini: Fooling Deep Structured Visual and Speech Recognition Models with Adversarial Examples, NIPS 2017

[13] Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, Raquel Urtasun, MultiNet: Real-time

Joint Semantic Reasoning for Autonomous Driving, arXiv:1612.07695

[14] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, The Cityscapes Dataset for Semantic Urban Scene Understanding, CVPR 2016.

[15] A Geiger, P Lenz, C Stiller, R Urtasun, Vision meets Robotics: The KITTI Dataset, International Journal of Robotics Research (IJRR) 2013

[16] Konda Reddy Mopuri, Aditya Ganeshan, R. Venkatesh Babu, Generalizable Data-free Objective for Crafting Universal Adversarial Perturbation, arXiv:1801.08092

[17] Clément Godard, Oisin Mac Aodha, Gabriel J. Brostow, Unsupervised monocular depth estimation with left-right consistency, CVPR , 2017

[18] Nicholas Carlini, David Wagner, Towards Evaluating the Robustness of Neural Networks, arXiv:1608.04644.

[19] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, Justin Gilmer, Adversarial Patch, arXiv:1712.09665.

[20] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, Patrick McDaniel, Ensemble Adversarial Training: Attacks and Defenses, ICLR 2018.