Adversarial Attack Against Scene Recognition System for Unmanned Vehicles

Xuankai Wang
1581025307@qq.com
Shanghai University of Electric Power
Shanghai, China

Mi Wen
wenmi2222@163.com
Shanghai University of Electric Power
Shanghai, China

Jinguo Li
lijg@shiep.edu.cn
Shanghai University of Electric Power
Shanghai, China

Zipeng Fu
fu-zipeng@engineering.ucla.edu
Department of Computer Science, University of California
Los Angeles, USA

Rongxing Lu
RLU1@unb.ca
University of New Brunswick
Canada

Kefei Chen
kfchen_shiep@163.com
Shanghai University of Electric Power
Shanghai, China

ABSTRACT

Unmanned scene recognition means that unmanned vehicles can collect environmental data from equipped sensors and make decisions through algorithms, in which deep learning has become one of key technologies. Especially, with the discovery of adversarial examples against deep learning, the research on offensive and defensive against adversarial examples illustrates that the deep learning model for unmanned scene recognition also has the safety vulnerability. However, as far as we know, few studies have tried to explore the adversarial example attack in this field. Therefore, we try to address this problem by generating adversarial examples againist scene recognition classification model through experiments. In addition, we also try to improve the adversarial model robustness by the adversarial training. Extensive experiments have been conducted, and experimental results show that adversarial examples have an efficient attack effect on the neural network for scene recognition.

KEYWORDS

adversarial examples, unmanned vehicle, scene recognition

ACM Reference Format:
Xuankai Wang, Mi Wen, Jinguo Li, Zipeng Fu, Rongxing Lu, and Kefei Chen. 2019. Adversarial Attack Against Scene Recognition System for Unmanned Vehicles. In *ACM Turing Celebration Conference - China (ACM TURC 2019) (ACM TURC 2019), May 17–19, 2019, Chengdu, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3321408.3321571

1 INTRODUCTION

As one of the most important means of transportation, the automobile occupies an important position in human civilization. With the development of economy, the total number of cars in the world is growing rapidly. This results in a series of traffic overload and traffic congestion problems. Thus, the inherent safety risks of automobiles are also increasing. According to the World Health Organization's Global Status Report on Road Safety, every year 1.25 million people die from car accident. Due to the accident, 35 million people are injured, and economic losses amount to hundreds of billions of dollars [1]. In this context, unmanned ground vehicles technology has becomes a hot research field. Unmanned vehicles can rely on the in-vehicle intelligent system to collect information on the surrounding environment in real time to make decisions. At the same time, it can rely on the communication system to exchange information with each other for the management of traffic. Therefore, the development of unmanned vehicles technology has great practical significance for alleviating traffic congestion and reducing traffic accidents.

Unmanned vehicles technology architecture is divided into three layers: a perception layer, a planning layer, and a control layer. The perception layer uses the camera, IMU and other sensors which are preset in the car to obtain the surrounding environment information, including the location of cars and pedestrians. The obtained results of perception layer will have a huge impact on the unmanned vehicle's driving behavior. Unmanned vehicles scene recognition is the main component of the perception layer. It refers to the use of in-vehicle equipment to identify environmental data and road information. Thus, the classification accuracy of scene recognition is very crucial to the safety of unmanned vehicles.

The main task of scene recognition technology is to identify and classify the scene images collected by the sensors. In recent years, deep learning has made many progress and achievements in the efficiency and accuracy of image recognition [2]. Meanwhile,convolutional neural networks [3] are particularly suit for classification tasks which have large data volume and rich input content. Accordingly, CNN and

ACM TURC 2019, May 17–19, 2019, Chengdu, China
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-7158-2/19/05...$15.00
https://doi.org/10.1145/3321408.3321571

related technologies have been widely used in the field of scene recognition. For example, Herranz et al. [4] proposed a CNN-obj model with a spliced structure to deal with scene recognition problems for instance.

However, as deep learning is susceptible to small disturbances [5], the potential risks of applying deep learning algorithms are gradually emerging. Adversarial example attack is that it misleads the result of the classification model by adding small disturbances to the legitimate examples. After the concept of adversarial examples was proposed [6] [7] [8] [9], it has become another hot research direction in the field of machine learning.In a short period, a variety of algorithms for generating adversarial examples have been proposed, and the successful generation rate on MNIST and CIRAF-10 has reached a very high level. We believe that adversarial example also has a huge potential hidden risk in scene recognition which uses image recognition as the core.As far as we know, few specific adversarial attack and defense research against unmanned scene recognition has been proposed.

In this paper, we try to use adversarial examples to attack the scene recognition classification model, and then use the adversarial training to improve the adversarial robustness of classification model to defense adversarial attack. The main contributions of this paper is shown as follows:

- First, we construct a deep learning neural network of unmanned scene recognition from the kitti dataset. Through multiple iterations, we get the classification accuracy of the classification model for legitimate examples. Then we use FGSM [6] and Deepfool [8] algorithms to generate adversarial examples separately, experiment and record the classification accuracy of the classification model on adversarial examples. The experimental results show that adversarial examples have a strong attack effect on the unmanned scene recognition classification model.
- Second, we use the method of adversarial training to improve the robustness of the scene recognition classification model. Then we test the classification accuracy of the original model and the adversarial training model against adversarial examples.

The rest of this paper is organized as follows: In Section II, we introduced the framework of the unmanned model and the role of scene recognition in it. Meanwhile,we have established a classification system model as the target of attack. The basic concepts and characteristics of adversarial example and the attack model are described in detail in Section III while the denfense model are described in Section IV. Then experimental evaluation and results analysis are performed in section V. Finally, we draw our conclusion in section VI.

2 SYSTEM MODEL

2.1 Unmanned Scene Recognition

The core of an unmanned system consists of three parts: perception, planning, and control. This system is essentially a layered structure in which perception, planning and control work at different layers but interacting with each other,as shown in Figure.1.

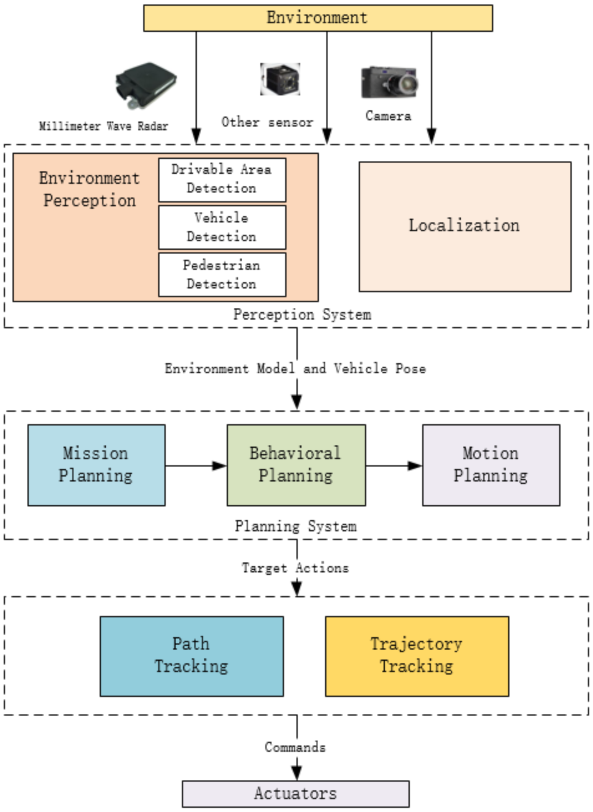

Figure 1: Unmanned vehicle system architecture

In order to ensure that the unmanned vehicle understands and grasps the environment, the Perception part of the unmanned system usually needs to obtain a large amount of information about the surrounding environment. Perception is the ability of an unmanned system to gather information from the environment and extract relevant knowledge from it. Environmental perception for unmanned vehicle refers to the semantic classification of the environment, such as the location of obstacles, the detection of road signs, and the detection of pedestrian vehicles. Unmanned vehicles usually acquire this information by combining data from various sensors such as Lidar, Camera, and Millimeter Wave Radar.Unmanned systems typically use image vision to perform road detection and detection of targets on the road. Road detection includes detection of road lines , detection of road signs on the road, detection of other vehicles, pedestrian detection, traffic signs and classification of all traffic participants. Among them, the detection and classification of traffic participants mainly rely on various deep learning models including convolutional neural networks. The classification model we have established for this study is a convolutional neural network for vehicle detection and pedestrian detection.

A scene is defined as a real-world environment which is semantically consistent and characterized by a namable human visual approach. Scene recognition is a technique for sifting images from similar images with similar scene semantic features and classifying all visual scenes. In general, scene recognition is based on the principle of human visual perception, extracting commonalities in image collections, and using these commonalities to complete the distinction between different scenes to achieve the purpose of scene recognition. Image scenes are often composed of multiple objects, while the spatial position and relationship layout formed by the combination of multiple objects are ever-changing. Therefore scene recognition pays more attention to the abovementioned layout relations than object recognition instead of focusing only on objects. In addition, the scene recognition is closer to the cognitive process of human vision, which is closer to abstract semantics and higher semantics than object recognition.

The specific application of scene recognition in the unmanned technology is also called road scene recognition. It means that the unmanned vehicle obtains the data information from the road through the sensor such as the on-board camera or radar during the road environment driving, then realizes the extraction of road topology geometric attributes and the identification of vehicle pedestrian position status in the road. The scene recognition work is essentially a classification task, so the scene recognition method can be abstracted into two phases as the classification method: feature extraction and training classifier. In recent years, the extraction of local features based on deep convolutional neural networks has become the mainstream method for scene feature extraction. Due to the strong learning ability of deep convolutional neural networks, it is possible to extract features of different abstraction levels layer by layer. Along with the improvement of GPU computing power and the establishment of large-scale data sets, the application of deep convolutional neural networks in scene recognition has been greatly improved.

2.2 Dataset

In this paper, the kitti dataset [9] is used to build the classification model and test the attack results. The KITTI Vision Benchmark Suite is a project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago.This dataset is generally used to evaluate the performance of computer vision technology, including image data collected from real scenes with varying degrees of occlusion. They used the improved Volkswagen Passat B6 as a platform to capture image data from urban, rural and other scenes and record them using computers and a real-time database. Sensors such as the Inertial Navigation System are used in the data collection process [10].

We used the test evaluation data for the Kitti dataset 2012. The object detection and object orientation estimation benchmark consists of 7481 training images and 7518 test images, comprising a total of 80.256 labeled objects. All images are color and saved as png. The original data set is divided into eight categories, 'Car', 'Van', 'Truck','Pedestrian', 'Person sitting', 'Cyclist', 'Tram','Misc' and 'DontCare'. Since there is no picture corresponding to the 'DontCare' and few picture in 'Person sitting' from the training set. We have included the image of the 'Person sitting' category into 'Pedestrian' and deleted the 'DontCare' class. The final integration into 'Car', 'Van', 'Truck', 'Pedestrian', 'Cyclist', 'Tram', 'Misc' 7 categories.

2.3 Classification Model

We use convolutional neural networks to generate classification models for unmanned scene recognition. CNN was first proposed by Yann LeCun and applied to handwritten font recognition (MINST) [3]. Convolutional neural networks are multi-layer neural networks that excel at dealing with related machine learning problems of images, especially large images. The input information is transmitted to different layers in turn, and the digital filter of each layer is used to obtain the most significant feature of the data. Since the local receptive field allows neurons to capture image underlying features such as image orientation edges and corner points, CNN's data transmission method can obtain significant features that are robust to scaling, rotation, and translation. Through a series of methods, the convolutional network has successfully addressed the image recognition problem with huge data volume and finally enabled it to be trained.

Convolutional neural networks are widely used in the field of computer vision. We use multicolor input data in our research and use RGB channels to process input images as shown in Figure 2. The hidden layer consists of a common construction of convolutional layer, pooling layer and fully-connected layer, using Rectified Linear Unit (ReLU) as the excitation function. The function of the convolutional layer is to extract the features of the input data. Convolution kernel size, step size and padding are important hyperparameters of the convolutional neural network, which jointly determine the size of the convolutional layer output feature map. After feature extraction in the convolutional layer, the output feature map is passed to the pooling layer for feature selection and information filtering. Then we use max pooling to process the data. Finally, the output is obtained by the fully-connected layer. The common convolutional neural network organization structure can be expressed as Figure 2.

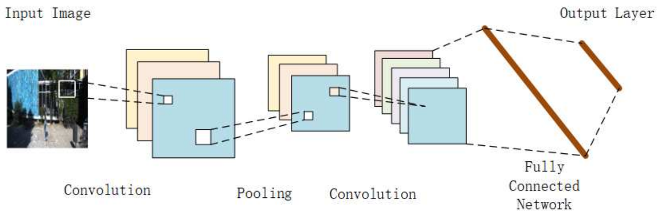

Figure 2: Classification system model(CNN)

3 ATTACK MODEL

In this section, we specifically introduced the conceptual characteristics of the adversarial example and detailed the attack model we used.

3.1 Adversarial example

The concept of adversarial examples was first proposed by Goodfellow et al [5]. When they studied the characteristics of neural networks, they found the results of neural network classification were susceptible to small disturbances. Adversarial example is a specially processed machine learning model input sample. Attacker adds a subtle distortion to the input sample in the data set, causing the classification model to give a false output with high confidence [6]. Most of the existing machine learning models, including neural network models, are susceptible to adversarial examples. Since the training set used to generate the classification model cannot cover all the possibilities, it is impossible to train a model that covers all the sample features, which leads to the inconsistency between the trained model boundaries and the real decision boundaries. Adversarial examples exist in the space thus generated [11].

We first generate a classification model F from a clean training set and give a legitimate input x. The correct label for sample x is y. If x' obtained by adding a small distortion r to the instance x can cause F to obtain the wrong classification result $F(x') \neq y$, we will refer to x' as an un-targeted adversarial example; If the attacker presets a target label $T \neq y$ and misleads the classification result of x' to $F(x') = T$, while the added distortion r is less than a certain threshold (the added distortion is small enough), we will refer to x' as a targeted adversarial example.

The amount of distortion is an important measure of the quality of adversarial example. The smaller the distortion of adversarial example, the closer it is to the original example, the more difficult it is to detect and recognize. Most adversarial examples attacks use an L_p distance to define closeness,defined as $\|v\|_p = \left(\sum_{i=1}^{n} |v_i|^p\right)^{\frac{1}{p}}$. There are three commonly used L_p distances: L_0 distance represents the number of pixels that change; L_2 distance measures the standard Euclidean distance between x and x'; L_∞distance measures the maximum change to any of the coordinates. Different attack algorithms may use different L_p distances to define closeness.

Adversarial attacks are mainly divided into two types, one is white box attack and the other is black box attack. The so-called white box attack refers to the attacker's full understanding of the target classifier, that is, the attacker knows the space where the classifier class tag is located, the type of the classifier, and the algorithm used by the machine learning; black box attack means that the attacker knows the feature representation of the target and the type of the classifier, but does not know the classifier form or training data to be learned.But the attacker can still interact with the machine learning system, for example, by entering any input to observe and judging the output of the classifier.

3.2 Attack Model

In the paper, we use FGSM and Deepfool [8] algorithms to generate adversarial examples against the unmanned scene recognition classification model. The attack model we built is a un-targeted attack based on the white box model, which means that the attacker can know the algorithm used by machine learning and the parameters used by the algorithm. The attacker can interact with the machine learning system in the process of generating adversarial attack data. The attack target is simply misleading the classifier to output an incorrect result, not a specified error label.

3.2.1 Fast gradient sign method. The Fast gradient sign method (FGSM) was proposed by Goodfellow in 2015 [6]. It is an algorithm based on gradient generation adversarial examples. The training objective is to maximize the loss function $J(x, y)$ to obtain the adversarial examples x', where J is the loss of the classification error in the classification algorithm function, usually taking the cross entropy loss. Maximizing J makes the examples after adding noise no longer judged by the classifier to belong to label y as shown in (2), thus achieving the purpose of attack. The L_∞ constraint $\|x' - x\|_\infty \leq \epsilon$ is required throughout the optimization process to limit the error between the adversarial example and the initial legitimate example to a certain extent.

$$x' = x + \epsilon \cdot sign\left(\bigtriangledown_x J(x, y)\right) \tag{2}$$

3.2.2 Deepfool. Deepfool [8] is an adversarial example generation method based on the idea of hyperplane classification. As we all know, in the two-class problem, the hyperplane is the basis for the classification. To change the classification of a sample x, the smallest disturbance is to move x to the hyperplane, which is the least costly. The problem of multiple classifications is similar.

Let us start with the linear binary classifiers problem and analyze the principle of the algorithm. We define $F \equiv \{x : f(x) = 0\}$ as the classification level, where f is a linear binary classifier $f(x) = w^T x + b$. To make x' into another labal, the size of the simplest perturbation is $\Delta\left(x', f\right)$ whose direction perpendicular to the classification level.

$$r_*(x') := argmin \left\|r\right\|_2 \tag{3}$$

$$subject\ to\ sign(f(x'+r)) \neq sign(f(x'))$$

$$= -\frac{f(x')}{\|w\|_2^2} w.$$

According to the formula (3), distortion r_* can be easily calculated, but this perturbation can only make the sample reach the classification surface, but not enough to pass, so the final perturbation is $r' = r_*(1+\eta), \eta \ll 1$,we take η as 0.02 in our experiment. In the same way of linear multi-classification problem, to find the minimum perturbation, only need to calculate the distance of x to multiple classification planes separately, and take the smallest one of them as the minimum perturbation.

4 DEFENSE MODEL

After attacking the unmanned scene recognition classification model, we try to use the adversarial training method to improve the robustness of the classification model and make a simple defense against the attack.

Adversarial training is the earliest proposed defense method against the sample. Its main idea is that in the model training process, the training samples are no longer just the original samples. The model adds the generated adversarial samples as new training samples to the training set. As a matter of course, as the model is trained more and more, the accuracy of the original image will increase on the one hand, and the robustness of the model against adversarial example will increase on the other hand.

We use the objective adversarial function as a specific solution to the training. After knowing the algorithm by which the attacker generates adversarial examples, we can construct the anti-target function for the attack mode to train the model [12].

5 EXPERIMENT

In this section, we will evaluate the attack effectiveness of adversarial example attack against the scene recognition neural network. The simulations are conducted on a 64 bit computer with dual Intel(R) Core(TM) i7-8750H 2.20 GHz CPU and 4G RAM by using Python, Numpy ,TensorflowCleverHans [13] and Keras to analyze data. CleverHans is a Python library to benchmark machine learning systems' vulnerability to adversarial examples.

We use the convolutional neural network to establish the classification model corresponding to kitti, set the number of epochs to train model to 6, the size of training batches to 128, and the learning rate of 0.001. At the same time, we use cross entropy as the loss function and take the amount of label smoothing for cross entropy to 0.1. By testing the trained classification model, we obtain the accuracy rate of the kitti classification model for the legitimate examples is 80.94%.

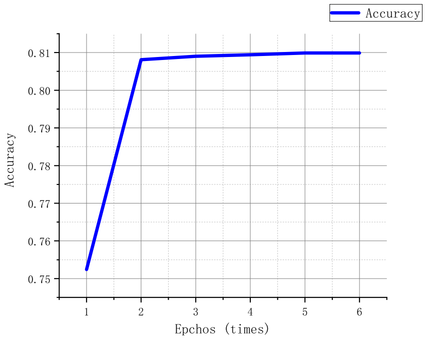

Figure 3: Accuracy rate during classification model training

Then we use the FGSM and Deepfool algorithms in Cleverhans to generate adversarial examples based on the kitti dataset, and use the classification model to classify the two types of adversarial examples. We performed ten experiments on each of the two attacks, the results are shown in the Figure 4, and the average of the accuracy is recorded in the table 1. According to the experiment, we found that adversarial examples have a good attack effect on the unmanned scene recognition. The adversarial attack generated by the FGSM algorithm is relatively weak in attack effect but also achieves an error classification rate of more than 80%, and the successful classification rate corresponding to adversarial examples of the Deepfool algorithm exceeds 90%. Considering the importance of scene recognition in the unmanned framework and the impact of misidentification on subsequent decision-making layers, unmanned driving poses a significant safety hazard in dealing with adversarial examples.

Table 1: Adversarial examples attack experiment result

	Legitimate	FGSM	Deepfool
Accuracy	80.94%	11.66%	6.49%

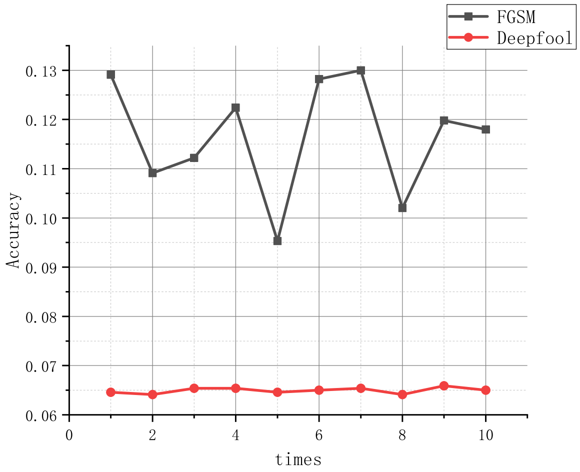

Figure 4: Adversarial examples attack experiment result

Finally, we use the adversarial training against FGSM algorithm to improve the robustness of the unmanned scene recognition classification model and test it again with adversarial examples generated by the FGSM algorithm. Experiments show that the adversarial training improves the defense effect of the classification model to a certain extent, and the successful attack rate of adversarial examples is reduced to 24.76%.

Table 2: Adversarial training experiment results

	Initial model	After adversarial training
Accuracy	11.66%	75.24%

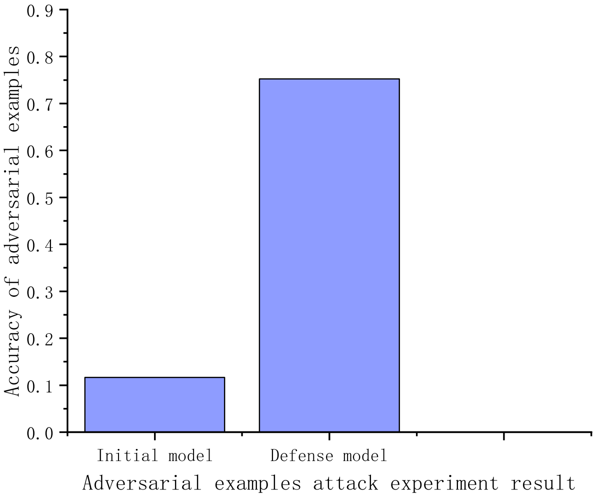

Figure 5: Adversarial examples attack experiment result

6 CONCLUSION

In this paper we propose that adversarial example attacks may pose risks and threats to the field of unmanned scene recognition. Then we conducted experimental evaluation through experiments, using FGSM and Deepfool algorithms to generate adversarial examples to attack the unmanned scene recognition classification model, and compared the attack effects of the two adversarial example generation algorithms. Finally, we conducted adversarial training against FGSM to improve the robustness of the classification model. For future work, on the one hand, we can replace the scene identification dataset with larger data volume or build a more complex scene recognition classification model as target model; on the other hand, we may try to find a better defense scheme based on the image features identified by the unmanned scene recognition.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under Grant No.61872230, No.61572311, No.61702321.

REFERENCES

[1] T Toroyan, M. M. Peden, and K Iaych. Who launches second global status report on road safety. *Injury Prevention Journal of the International Society for Child & Adolescent Injury Prevention*, 19(2):150–150, 2013.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *International Conference on Neural Information Processing Systems*, pages 1097–1105, 2012.

[3] S. Haykin and B. Kosko. Gradientbased learning applied to document recognition. In *IEEE*, pages 306–351, 2009.

[4] Luis Herranz, Shuqiang Jiang, and Xiangyang Li. Scene recognition with cnns: objects, scales and dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 571–579, 2016.

[5] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *Computer Science*, 2013.

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples (2014). *arXiv preprint arXiv:1412.6572*.

[7] Nicolas Papernot, Patrick Mcdaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy*, pages 372–387, 2016.

[8] Seyed Mohsen Moosavidezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Computer Vision and Pattern Recognition*, pages 2574–2582, 2016.

[9] Andreas Geiger. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[10] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[11] Patrick Mcdaniel, Nicolas Papernot, and Z. Berkay Celik. Machine learning in adversarial settings. *IEEE Security & Privacy*, 14(3):68–72, 2016.

[12] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[13] Nicolas Papernot, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Fartash Faghri, Alexander Matyasko, Karen Hambardzumyan, Yi-Lin Juang, Alexey Kurakin, Ryan Sheatsley, et al. cleverhans v2. 0.0: an adversarial machine learning library. *arXiv preprint arXiv:1610.00768*, 2016.