Deep learning systems have achieved great success in various types of applications over recent years. They are increasingly being adopted for safety critical tasks, such as face recognition, surveillance systems, speech recognition, and autonomous driving. On the other hand, it has been found that deep neural networks (DNNs) can be easily fooled by adversarial input samples. These imperceptible perturbations on images can lead any machine learning system to misclassify the objects with high confidence. Furthermore, they can be almost indistinguishable to a human observer. These systems can also be exposed to adverse weather conditions such as fog, rain, and snow. This vulnerability raises major concerns in security-sensitive environments. Therefore, vulnerability of deep learning systems to synthetic adversarial attacks has been extensively studied and demonstrated, but the impact of natural weather conditions on these systems has not been studied in detail.

The main contribution of this thesis is exploring the effects of fog on classification accuracy of the popular Inception deep learning model. We use stereo images from the Cityscapes dataset and computer graphics techniques to mimic realistic naturally occurring fog. We show that the Inception deep learning model is vulnerable to the addition of fog in images.

We also review the types of adversarial attacks and defenses, describe the state-of-the-art methods for each group, and compare their results.

Adversarial images can be used to generate targeted attacks or non-targeted attacks. Targeted attacks misguide the deep learning networks to produce responses from a specific a priori determined class. In non-targeted attacks, all images in the dataset are not assigned to a specific class; instead, the output of the deep neural network is arbitrarily wrong. In this thesis, we create non-targeted, iterative, and physical attacks.

Approved for distribution by Sumit Kumar Jha, Committee Chair, on October 1, 2019.