# ABSTRACT

Artificial intelligence is already used in many applications, such as voice assistants or driver assistance systems. In many cases, such systems are based on artificial neural networks. In the field of image classification Convolutional Neural Networks (CNNs) are state of the art models, achieving higher a accuracy than humans. However, these models are vulnerable to adversarial images. Adversarial Images are natural images which have been modified by an adversarial perturbation specifically designed to fool the classifier. The added pertubations are regularly undetectable for humans. Due to this vulnerability, the use of CNNs in safety-critial systems should be questioned critically.

There already exist methods which try to protect the models in different ways. Some methods modify the networks, transform the input images or train a detector. However, so far there is no defence method which is robust to all known and unknown attacks. In order to further advance research in this field, it is necessary to reproduce and compare the existing approaches and results. But in this context there are several limitations which inhibit the reasearch in this field, such as the use of different data sets, evaluation metrics, frameworks or unavailable code.

This work gives an insight into the theoretical aspects of Adversarial Images, by presenting, among other things, methods for generating and defending against them. Furthermore, the work shows the realization of a benchmark of reactive defense methods, which try to solve the mentioned limitations in order to compare various methods. The benchmark is based on a uniform data set of Adversarial Images created in this thesis. The benchmark is applied to defense methods that try to eliminate the adversarial pertubation by image transformations.This thesis shows that these reactive methods represent a very simple and cost-effective approach, but do not show the positive effects on the generated adversarial images promised in the respective papers.



**Keywords:** Machine Learning, Convolutional Neural Networks, Adversarial Images, Adversarial Examples

# ZUSAMMENFASSUNG

Künstliche Intelligenz findet immer mehr Einzug in das alltägliche Leben, z.B. in Form von Sprachassistenten oder Fahrerassistenzsystemen. In vielen Fällen basieren diese Systeme auf künstlichen neuronalen Netzen. Eine Art dieser Modelle sind Convolutional Neural Networks (CNNs), die u.a. in der Bildklassifikation eingesetzt werden. Im Jahr 2013 entdeckten Forscher erstmals, dass CNNs eine Instabilität gegenüber gezielt erstellten Störungen aufweisen, die einem Eingabebild hinzugefügt werden, sogenannter Adversarial Images. Während derartige manipulative Störungen für den Menschen i.d.R. nicht wahrnehmbar sind, so führen sie jedoch zu einer fehlerhaften Klassifikation des CNNs. Aufgrund dieser Verletzlichkeit ist der Einsatz solcher Modelle in sicherheitskritischen Systemen kritisch zu hinterfragen.

Es existieren bereits unterschiedliche Methoden, welche durch Modifikation des CNNs, eine Transformation des Eingabebildes oder der Verwendung eines Dektektors versuchen, das Modell vor Angriffen mit Adversarial Images zu schützen. Allerdings existiert bisher keine Abwehrmethode, die ein Modell gegen alle bekannten und unbekannten absichern kann. Um die Forschung in diesem Gebiet weiter voranzutreiben ist es hilfreich die existierenden Ansätze und Ergebnisse reproduzieren und vergleichen zu können. In diesem Kontext bestehen jedoch einige Schwierigkeiten, bspw. unterschiedliche verwendete Datensätzen und Bewertungsmetriken oder nicht öffentlich zugänglicher Code.

Diese Arbeit gibt einen Einblick in die theoretischen Facetten von Adversarial Images, indem u.a. einige Methoden zur Generierung und Verteidigung vorgestellt werden. Darüber hinaus zeigt die Arbeit die Durchführung eines Benchmarks von reaktiven Abwehrmethoden, um einen Vergleich verschiedener Methoden zu ermöglichen und auf diese Weise die beschriebenen Schwierigkeiten zu beheben. Der Benchmark findet dabei auf einem einheitlichen und in dieser Arbeit erstellten Datensatz von Adversarial Images statt. In diesem praktischen Teil wurden Abwehrmethoden untersucht, die mittels Bildtransformationen versuchen, die manipulativen Störungen unschädlich zu machen. Die Ansätze werden dabei auf drei verschiedenen CNNs durchgeführt und in mehreren Szenarien getestet sowie evaluiert. Die Masterarbeit zeigt, dass reaktiven Methoden einen sehr einfachen und kostengünstigen Ansatz darstellen, jedoch nicht die in den jeweiligen Papern versprochenen positiven Effekte auf den generierten Adversarial Images zeigen.

**Schlagwörter:** Maschinelles Lernen, Convolutional Neural Networks, Adversarial Images, Adversarial Examples