Securing Deep Learning against Adversarial Attacks for Connected and Automated Vehicles

Technology Transfer Activities

by

Pierluigi Pisu, Ph.D., Clemson University
Gurcan Comert, Ph.D., Benedict College
Negash Begashaw, Ph.D., Benedict College
Chunheng Zhao, Ph.D. student, Clemson University

Contact information

Pierluigi Pisu, Ph.D.
4 Research Drive, Greenville, SC 29607
Clemson University
Phone: (864) 283-7227; *E-mail:* pisup@clemson.edu

April 2023

Center for Connected Multimodal Mobility (C^2M^2)

CLEMSON UNIVERSITY

Benedict College | The Citadel, The Military College of South Carolina | SC State University | University of South Carolina

200 Lowry Hall, Clemson University
Clemson, SC 29634

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by the Center for Connected Multimodal Mobility (C^2M^2) (Tier 1 University Transportation Center) Grant, which is headquartered at Clemson University, Clemson, South Carolina, USA, from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

ACKNOWLEDGMENT

Insert your text here.

Table of Contents

TECHNOLOGY TRANSFER ACTIVITIES

1 Outputs

This project proposes a deep ensemble model for image classification with the fusion of discriminative features and generative models. A causal graph is designed while constructing the deep Bayes classifier to model the adversarial perturbations. As the deep Bayes classifier can't achieve the state-of-the-art accuracy of discriminative classifiers, we fuse the object features extracted from pre-trained CNNs with original images as final inputs. Benefiting from this structure, the proposed method is generic and can be applied to various discriminative classifiers. The generative model can be used as an auxiliary network to be built on top of any pre-trained CNNs. Experimental results show that the proposed ensemble model achieves reduced accuracy loss against adversarial examples and gains better overall model causality and interpretability.

1.1 Output #1

Dissemination of C^2M^2 research results was obtained through publications, workshop, webinar, and technical report as indicated below:

C. Zhao, P. Pisu, G. Comert, V. Vaidyan, N. Begashaw and N. C. Hubig, “A Robust Adversarial Ensemble with Causal (Feature Interaction) Interpretations for Image Classification”, IEEEE/RSJ International Conference on Intelligent Robots and Systems, 2023, submitted for review.

“Securing Deep Learning against Adversarial Attacks for Connected and Automated Vehicles”, Doctoral Symposium on 14th Annual Conference of the Prognostics and Health Management Society, Oct. 31st, 2022.

P. Pisu, G. Comert, N. Begashaw and C. Zhao, “Securing Deep Learning against Adversarial Attacks for Connected and Automated Vehicles”, Final Technical Report, C^2M^2, Apr. 2023.

Technical Reports: 1; Papers: 1, Presentations: 1

1.1 Output #2

The following new methods and products resulted from this research:

- A bottom-up discriminative-generative ensemble model for image classification is developed, which leverages both generative and discriminative models with built-in adversarial causal relationships. A causal graph with latent variables is created to build Bayes-based generative classifier. The inputs consist of both original inputs and discriminative features.
- An evaluation method using adversarial examples as counterfactual metrics is proposed. The proposed ensemble model not only shows better classification accuracy against adversarial examples but also shows better model causality.

New Research Products: 2

1.1 Output #3

At this time, no demonstration is done for the proposed method.

2 Outcomes

The project allowed for collaboration of faculty from Clemson University and Benedict College and also training of students at graduate and undergraduate level working on the project.

2.1 Outcome #1

1. This project facilitated 1 more proposal.

2. One graduate student (Clemson University) and four undergraduate students at Benedict College were trained.

3. Three additional (1 mathematics, 1 computer engineering, and 1 electrical engineering) faculty collaborated from Benedict College.

4. Three undergraduate students were continuously involved in the project and transitioned into relevant careers. One student (May 2022 graduate) accepted an internship at IBM, New York, and PhD study at Harvard University in computer science. One student (May 2023 graduate) got accepted to UCLA's PhD program in mechanical engineering and MS program at Clemson University in civil engineering. Another student (May 2023 graduate) accepted a computational laboratory engineering position at Benedict College.

2.2 Outcome #2

We leverage Clemson Palmetto Computing platform for the use of cluster computing and high-performance computing. The incorporation of the proposed deep neural network algorithm and existing Cluster computing platforms is conducted in this project. We leverage F1/10th RCcar testbed for the experiments and Nvidia Jetson for the use of embedded computing in this project.

2 Outcome #3

This proposal generated robust methods for object detection and classification in perception module on connected and automated vehicles. It will help reduce the misclassifications of surrounding environments in case of adversarial attacks from the vehicle on-board sensors, thus, be able to improve CAV security and help make correct path and behavior plannings.

3 Impacts

This project focuses on the development of new tools for making DNNs more resilient to adversarial attacks with particular focus on the development of object detectors utilized in the perception module of CAVs. The results of this project will have a broad applicability not only to the transportation sector but also to many other engineering fields such as autonomous driving, biometric identification, speech and face recognition, intelligent transportation systems, and robotics (vision SLAM).

3.1 Impact #1

At this time, we are not aware of any adoption by transportation agencies.

3.2 Impact #2

Currently, our method was tested on the image dataset CIFAR-100 and showed resiliency against adversarial images, which can significantly increase the classification accuracy to help reduce the risk of misclassification and improve safety.