Evaluation of Robustness of Off-Road Autonomous Driving Segmentation against Adversarial Attacks: A Dataset-Centric Study

Extended Abstract

Pankaj Deoli
Robotics Research Lab, RPTU
Kaiserslautern, Germany
deoli@cs.uni-kl.de

Rohit Kumar
Robotics Research Lab, RPTU
Kaiserslautern, Germany
r_kumar22@cs.uni-kl.de

Axel Vierling
Robotics Research Lab, RPTU
Kaiserslautern, Germany
vierling@cs.uni-kl.de

Karsten Berns
Robotics Research Lab, RPTU
Kaiserslautern, Germany
berns@cs.uni-kl.de

ABSTRACT

The study explores the vulnerability of semantic segmentation models to adversarial input perturbations in the domain of off-road autonomous driving. Existing studies have primarily concentrated on enhancing model's robustness via architectural modifications along-with using noisy images during training. On the contrary, little attention has been paid to investigating the impact of datasets on the adversarial attacks. Our study aims to address this gap by examining the impact of non-robust features in off-road datasets and comparing the effects of adversarial attacks on different segmentation network architectures. To enable this, a robust dataset is created consisting of only robust features and training the networks on this robustified dataset. We present both qualitative and quantitative analysis of our findings. The code is publicly available at ***https://github.com/rohtkumar/adversarial_attacks_on_segmentation***

KEYWORDS

Adversarial examples; Off-road autonomous driving; Semantic segmentation; Robotics

ACM Reference Format:
Pankaj Deoli, Rohit Kumar, Axel Vierling, and Karsten Berns. 2024. Evaluation of Robustness of Off-Road Autonomous Driving Segmentation against Adversarial Attacks: A Dataset-Centric Study: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

With the advent of high-stakes applications, such as autonomous driving, which heavily relies on Deep Neural Networks (DNNs) for perception, people are increasingly fascinated by the capabilities of machines. However, safety concerns have arisen in relation to autonomous driving with respect to DNNs, particularly due to the vulnerability of such systems to adversarial attacks. Significant research and developmental efforts have been carried out in the domain of Adversarial examples for general classification tasks but little for off-road autonomous driving. Further, these improvements have been inclined towards urban contexts [13], [19] since addressing man-made structural entities is relatively simple. On the other hand, great effort has been put towards segmenting paths, trees, and plants in off-road areas, but little towards enhancing the robustness of autonomous driving systems to adversarial examples in such environments. This work addresses this gap by evaluating the effects of adversarial attacks for semantic segmentation in the context of off-road autonomous driving. For this, an off-road robust dataset has been created (by adapting [7]) consisting of only robust features. Along-with this, an analysis of this robustified dataset on two State-of-the-art (SOTA) semantic segmentation networks, the their comparative analysis is provided.

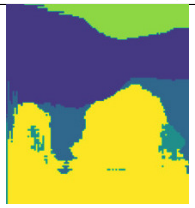
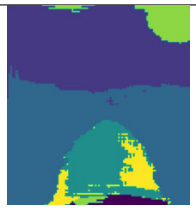
This is an extended abstract. The full version of this paper can be found online [4].

2 APPROACH

Various gradient-based adversarial methods such as *Fast Gradient Sign Method (FGSM)* [6], *Basic Iterative Method (BIM)* [8] and *Projected Gradient Descent (PGD)* [8] exploit the use of gradients in ML models to generate examples. Among them, PGD is more effective against well-trained models since it includes a projection step to ensure that the perturbed input data remains within a certain range. And therefore, it was selected as the main attack (with L^2 and L^∞ bounds) for our task at hand. Four baseline semantic segmentation networks FCN16 [10], UNet [12], DeepLab [2], PNPNet [9] and LinkNet [1] were explored. Because of the robustness of UNet (to work with fewer training images to produce precise segmentation maps) and capability of LinkNet (to recover lost spatial information) along-with good performance with less parameters, these 2 were chosen as the baseline models. When compared to widely used urban datasets for autonomous driving such as KITTI (12919 images) [5], Cityscapes [3] (2500 images), Waymo [15] (1000 images),

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

Table 1: Networks performance on the robustified dataset. *R* represents robustified dataset along-with the respective adversarial attacks. Networks struggle to capture the details of an ambiguous environment.

Input image	Ground truth	UNet (RPGD L^2)	UNet (RPGD L^∞)	LinkNet (RPGD L^2)	LinkNet (RPGD L^∞)
					
					

off-road datasets lack distinctiveness and quantity. For this reason, Freiburg forest [16] and Yamaha CMU Off-road dataset (YCOR) [11] were combined (a total of 1442 images) with final 10 classes as Background, Vegetation, Traversable, grass, Smooth, trail, Obstacle, Sky, Rough, trail, Puddle, Non-Traversable Vegetation and Tree.

3 EXPERIMENTATION

To have a systematic analysis of the effect of non-robust features in off-road environment, the experiments were divided into 4 stages i.e.

- **Standard training on the merged dataset** : When both UNet and LinkNet were trained for 100 epochs with the configurations mentioned in ([4] (table 2)), UNet achieved an IoU of 73% and 69% during training and evaluation respectively. Similarly, LinkNet achieved an IoU of 71% and 69% during training and evaluation respectively ([4] (table 3)).
- **Adversarial training (*PGD* (L^2 *and* L^∞))** : When UNet was trained on PGD L^2 and L^∞ individually, the network showed good performance quantitatively but lacked precision during separating different surface types. When LinkNet was trained with the same configurations, it was found that when training on broad attack families (i.e. training on PGD L^∞), the test loss decreased considerably.
- **Robustifying the merged dataset** : The combined dataset was robustified following the approach mentioned in [7] by separating the robust features from non-robust features.
- **Standard training on Robustified dataset** : When both the adversarially trained networks were trained in the robustified dataset with the standard training configurations, the quantitative results were considerably improved. However, the qualitative results showcased poor model convergence and improper predictions thereby having contradicting results as seen with table 1.

4 FINDINGS

The work explored whether the dataset has any effect in countering the adversarial effect on semantic segmentation and whether it can be changed (based on attacks) and still maintain the same performance w.r.t state of the art metrics. The study done by [7] depicts that the robustified model performs better, however, the qualitative results were not talked about since it focused on classification rather than segmentation. Some factors included :

- **Presence of Multi-class in input images** : The previous study by [7], in this regard, concentrated on binary classification. However, in our study, the presence of ambiguous (non-rich features) multi-class of the input space can present fundamental barriers to classifier's robustness. A classifier cannot be resistant against tiny perturbations since, at the highest level, for certain data distributions, any decision boundary will be near a significant portion of inputs.
- **Insufficient data** : Another major problem was the unavailability of a huge, distinct dataset with higher variability. Authors of [14] present that for appropriate learning, a good robust classifier requires ($O\sqrt{d}$) samples (d being the dimentionality of the data). In this paradigm, adversarial examples appear as a result of insufficient knowledge of the real data distribution. In particular, since training models robustly reduce the effective amount of information in the training data, more samples should be required to generalize robustly.

5 CONCLUSION

The work explored the role of datasets for achieving the segmentation robustness in off-road environments. The work presented a method to generalize the classifier's robustness, the analysis depict the unreliable correlation between qualitative and quantitative results. Efficient methodological transfer to unmanned robots such as Unimog requires good perception robustness [18], [17]. Future works still requires efficient, robust perception strategies for safe navigation in rough off-road environments.

REFERENCES

[1] Abhishek Chaurasia and Eugenio Culurciello. 2017. LinkNet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, St. Petersburg, FL, USA, 1–4. https://doi.org/10.1109/vcip.2017.8305148

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 2017. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. arXiv:1606.00915 [cs.CV]

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. arXiv:1604.01685 [cs.CV]

[4] Pankaj Deoli, Rohit Kumar, Axel Vierling, and Karsten Berns. 2024. Evaluating the Robustness of Off-Road Autonomous Driving Segmentation against Adversarial Attacks: A Dataset-Centric analysis. arXiv:2402.02154 [cs.CV]

[5] A Geiger, P Lenz, C Stiller, and R Urtasun. 2013. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research* 32, 11 (2013), 1231–1237. https://doi.org/10.1177/0278364913491297 arXiv:https://doi.org/10.1177/0278364913491297

[6] Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. 2017. Adversarial Attacks on Neural Network Policies. arXiv:1702.02284 [cs.LG]

[7] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. Adversarial Examples Are Not Bugs, They Are Features. arXiv:1905.02175 [stat.ML]

[8] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2017. Adversarial examples in the physical world. arXiv:1607.02533 [cs.CV]

[9] Ming Liang, Bin Yang, Wenyuan Zeng, Yun Chen, Rui Hu, Sergio Casas, and Raquel Urtasun. 2020. PnPNet: End-to-End Perception and Prediction with Tracking in the Loop. arXiv:2005.14711 [cs.CV]

[10] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully Convolutional Networks for Semantic Segmentation. arXiv:1411.4038 [cs.CV]

[11] Daniel Maturana, Po-Wei Chou, Masashi Uenoyama, and Sebastian Scherer. 2018. Real-Time Semantic Mapping for Autonomous Off-Road Navigation. In *Field and Service Robotics*, Marco Hutter and Roland Siegwart (Eds.). Springer International Publishing, Cham, 335–350.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597 [cs.CV]

[13] Giulio Rossolini, Federico Nesti, Gianluca D'Amico, Saasha Nair, Alessandro Biondi, and Giorgio Buttazzo. 2022. On the Real-World Adversarial Robustness of Real-Time Semantic Segmentation Models for Autonomous Driving. arXiv:2201.01850 [cs.CV]

[14] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Mądry. 2018. Adversarially Robust Generalization Requires More Data. arXiv:1804.11285 [cs.LG]

[15] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Sheng Zhao, Shuyang Cheng, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. 2020. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. arXiv:1912.04838 [cs.CV]

[16] Abhinav Valada, Gabriel Oliveira, Thomas Brox, and Wolfram Burgard. 2017. Deep Multispectral Semantic Scene Understanding of Forested Environments Using Multimodal Fusion. In *Proceedings in Advanced Robotics, vol 1*. Springer, Cham, cham, 465–477. https://doi.org/10.1007/978-3-319-50115-4_41

[17] Patrick Wolf and Karsten Berns. 2021. Data-fusion for robust off-road perception considering data quality of uncertain sensors. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Prague, 6876–6883. https://doi.org/10.1109/IROS51168.2021.9636541

[18] Patrick Wolf, Pankaj Deoli, Satish Kumar Thangellapally, and Karsten Berns. 2023. Traction Optimization for Robust Navigation in Unstructured Environments Using Deep Neural Networks on the Example of the Off-Road Truck Unimog. In *Intelligent Autonomous Systems 17*, Ivan Petrovic, Emanuele Menegatti, and Ivan Marković (Eds.). Springer Nature Switzerland, Cham, 561–579.

[19] Maksym Yatsura, Kaspar Sakmann, N. Grace Hua, Matthias Hein, and Jan Hendrik Metzen. 2023. Certified Defences Against Adversarial Patch Attacks on Semantic Segmentation. arXiv:2209.05980 [cs.CV]