Adversarial Training and Robustness in Machine Learning Frameworks

[1]Mrs. Sangeetha G, [2]Mr. Bharath K, [3]Mr. Balamanikandan S, [4]Mr. Bharath G

Department of Computer Science[1,2,3,4]

SRM Valliammai Engineering College, Chennai, Tamil Nadu, India

[1]sangeethag.cse@srmvalliammai.ac.in, [2]bharathkannan.b47@gmail.com

[3]balamanikandanseenivasan@gmail.com, [4]bharath03112001@gmail.com

Abstract*: In the realm of machine learning, ensuring robustness against adversarial attacks is increasingly crucial. Adversarial training has emerged as a prominent strategy to fortify models against such vulnerabilities. This project provides a comprehensive overview of adversarial training and its pivotal role in bolstering the resilience of machine learning frameworks. We delve into the foundational principles of adversarial training, elucidating its underlying mechanisms and theoretical underpinnings. Furthermore, we survey state-of-the-art methodologies and techniques utilized in adversarial training, encompassing adversarial example generation and training methodologies. Through a thorough examination of recent advancements and empirical findings, we evaluate the effectiveness of adversarial training in enhancing the robustness of machine learning models across diverse domains and applications. Additionally, we address challenges and identify open research avenues in this burgeoning field, laying the groundwork for future developments aimed at strengthening the security and dependability of machine learning systems in real-world scenarios. By elucidating the intricacies of adversarial training and its implications for robust machine learning, this paper contributes to advancing the understanding and application of techniques crucial for safeguarding against adversarial threats in the evolving landscape of artificial intelligence.*

Keywords: Adversarial Training, Robustness, SGD, Model enhancement

REFERENCES

[1]. Zhang, H., et al. (2019). Theoretically principled trade-off between robustness and accuracy. In International Conference on Learning Representations.

[2]. Athalye, A., et al. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In International Conference on Machine Learning.

[3]. Goodfellow, I., et al. (2017). The limitations of deep learning in adversarial settings. In IEEE European Symposium on Security and Privacy.

[4]. Papernot, N., et al. (2017). Practical black-box attacks against machine learning. In Asia Conference on Computer and Communications Security.

[5]. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In IEEE Symposium on Security and Privacy.

[6]. Papernot, N., et al. (2016). Towards the science of security and privacy in machine learning. In Workshop on Artificial Intelligence and Security.

[7]. Szegedy, C., et al. (2014). Intriguing properties of neural networks. In International Conference on Learning Representations.

[8]. Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial examples in the physical world. In International Conference on Learning Representations.

[9]. Madry, A., et al. (2018). Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations.

DOI: 10.48175/IJARSCT-15935

[10]. Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In International Conference on Learning Representations.

DOI: 10.48175/IJARSCT-15935