When Digits Lie: An Analytical Study of Adversarial Attacks on Digit Recognition

Anumaneni Venkat Balachandra*
Department of CSE
R.V. College of Engineering
Bengaluru, India
anumanenivb.cd22@rvce.edu.in

Joel Stephen Mathew*
Department of CSE
R.V. College of Engineering
Bengaluru, India
jstephenmathew.cy22@rvce.edu.in

Aditya Suresh Nair
Department of CSE
R.V. College of Engineering
Bengaluru, India
adityasnair.cy22@rvce.edu.in

Arman Singh Bhati
Department of CSE
R.V. College of Engineering
Bengaluru, India
armansinghb.cy22@rvce.edu.in

Mohana
Department of CSE
R.V. College of Engineering
Bengaluru, India
mohana@rvce.edu.in

***Abstract*—- Briefly state the problem your paper addresses. - Summarize the methodology used. - Highlight key results. - Explain the significance of the findings. - Limit the abstract to 150-250 words.**

***Index Terms*—Select 4-6 keywords from the IEEE Thesaurus for discoverability.**

I. Introduction

A. Problem Statement

Clearly articulate the problem your paper aims to solve.

B. Background and Motivation

- Discuss the historical context of character recognition, highlighting the evolution of techniques from manual to automated systems. - Emphasize the importance of adversarial attacks on neural networks in the modern context.
(Find 2 papers)

C. Importance of Character recognition

Explain why accurate, automated handwritten digit recognition is crucial, providing practical examples and citing 2-3 relevant papers.

D. Overview of Handwritten Digit Classification using NN

Provide an overview of the neural network architecture used for digit classification, focusing on the parts susceptible to adversarial attacks. Use block diagram is possible and cite at least 1-2 papers

E. Objectives and Scope

Clearly define the objectives of your study. - Justify the choice of the MNIST dataset as a representative sample for real-world handwritten material. - Outline the scope of your research, including the specific adversarial techniques analyzed.

*Equal Contribution

II. Literature Review

A. Survery of Adversarial Techniques

- Provide detailed descriptions of the four adversarial techniques used in your study. Include methodology, application, and limitations. - Cite recent and relevant IEEE papers to support your discussion.(1-2 papers for each technique)

Bonus: find a paper that includes these techniques as the most popular adversarial techniques out there.

III. Methodology

A. Data Collection

Ran each combination of (attack method, original image, intended adversarial image) 50 times. Recorded the averages of the original prediction probability and new prediction probability and the most common adversarial image for each combination.

This generated a lot of data, which we condensed into 4 attack methods $\times$ 10 original images $\times$ 10 adversarial images.

- Describe the experimental setup, including the selection of MNIST images and the application of adversarial techniques. - Justify the rationale for running each combination 50 times and explain how this impacts the reliability of your results.

B. Data Analysis

Use Colab notebook. To start with, forget we have 4 methods and only focus on FGSM. Figure out the structure of our analysis and which graphs go in which order and how do we present the metrics and inferences.

Whatever template we come up with for FGSM, we copy paste for remaining methods quickly.

- Detail the steps taken to analyze the data, starting with FGSM as a template. - Explain the structure of the analysis, the order of presenting graphs, metrics, and inferences. - Ensure consistency in the analysis across all adversarial techniques.

IV. Results

A. Analysis of Results

- Present your findings using appropriate graphs, tables, and charts. Ensure clear captions and labels for all figures. - Discuss the effectiveness of each adversarial technique based on the collected data.

B. Comparison of Techniques

- Compare the performance of the adversarial techniques on various metrics such as accuracy degradation and computational cost. - Highlight significant patterns, trends, and insights from the data.

V. Discussion

A. Interpretation of Results

- Provide a detailed interpretation of the results, explaining what they mean in the context of adversarial machine learning.

B. Implications for ML security

- Discuss how the findings can be applied to improve the security and robustness of machine learning models in real-world scenarios.

C. Limitations

- Acknowledge the limitations of your study, such as dataset constraints, model assumptions, or the generalizability of the findings.

VI. Applications

A. Relevance to Real World Applications

- Identify five areas where your findings can be applied to enhance machine learning model security (e.g., finance, healthcare, autonomous vehicles). - Support each application with relevant literature, even if the papers do not directly relate to handwritten digit recognition.

VII. Future Work

A. Exapansion of Research

- Discuss the potential for exploring additional adversarial techniques, using different datasets, or applying the findings to other domains.

B. Potential Improvements

Suggest how the current methodology could be refined or expanded, such as by incorporating larger datasets or more sophisticated analysis techniques.

VIII. Conclusion

A. Summary of Findings

Recap the key takeaways from your research, emphasizing the contribution to the field of adversarial machine learning.

B. Impact

Reinforce the significance of your findings in advancing the understanding and security of machine learning models.

References

[1] V. Manieniyan, M. Thambidurai, and R. Selvakumar, "Study on energy crisis and the future of fossil fuels," in *Proceedings of SHEE*, vol. 10, pp. 2234-3689, 2009.