BTP Report

Manurbhav Arya

CS21B057

Guide: Dr. Chalavadi Vishnu

Indian Institute of Technology, Tirupati

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission to the best of my knowledge. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

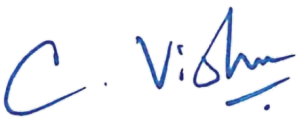
Place: Tirupati
Date: 09-05-2025

Signature
Manurbhav Arya
CS21B057

Note: If more than one authors mentioned in the cover page, modify each 'I' by 'we' and then include remaining author's name at the bottom of the page. Signature of each author in this page is must.

Certificate

This is to certify that the thesis titled **Adversarial Robustness in Bayesian Neural Networks for Real-World Systems**, submitted by **Manurbhav Arya (CS21B057)**, to the Indian Institute of Technology Tirupati, for the award of the degree of **Bachelor of Technology** , is a bonafide record of the research work done by him under my supervision. The contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.



Place: Tirupati
Date: 09-05-2025

Dr. Chalavadi Vishnu
Guide
Assistant Professor
Department of Computer Science and Engineering
IIT Tirupati - 517619

Note: If supervised by more than one professor, each professor's name must be included and signatures taken from all supervisors. Change "him" to "them" and "my" to "our" as appropriate.

Acknowledgments

I sincerely thank Dr. Chalavadi Vishnu for his constant guidance, encouragement, and support throughout this project. His expertise and advice played a crucial role in shaping our research and helped us navigate the challenges we faced. I truly value his mentorship, which not only deepened my technical knowledge but also motivated me to strive for excellence. Thank you for being an essential part of the successful completion of this project.

Contents

Adversarial Robustness in Bayesian Neural Networks for Real-World Systems

Abstract

Deep neural networks (DNNs) play a crucial role in many modern systems but remain vulnerable to adversarial attacks, posing risks in critical applications. Bayesian Neural Networks (BNNs) offer improved robustness through uncertainty estimation, yet they are not fully resistant to adversarial threats. This report summarizes previous work on adversarial defense, including BNN defense testing on CIFAR-10, and explores strategies to improve security in real-world deployments.

1 Introduction

1.1 Motivation

Machine learning models are being increasingly adopted across diverse domains such as healthcare, finance, autonomous systems, and industrial automation. These models are susceptible to adversarial attacks that can lead to degraded performance and unsafe outcomes. Ensuring the robustness of such systems is essential for reliable and secure deployment, particularly in safety- and mission-critical environments.

1.2 Problem Statement

While Bayesian Neural Networks (BNNs) provide uncertainty-aware predictions that enhance resilience, they still face adversarial vulnerabilities. Existing defense techniques, including adversarial training, offer partial protection but are not foolproof. This work investigates methods to further improve adversarial robustness using probabilistic deep learning techniques for general-purpose systems.

2 Previous BTP Work

2.1 Objective

The research explores the robustness of Bayesian Neural Networks (BNNs) against adversarial attacks, focusing on how adversarial noise affects the distribution of neural network nodes modeled with defined mean and variance. The study investigates the impact of adversarial attacks on these distributions and evaluates methods to control and mitigate these effects.

2.2 Methodology

2.2.1 Model Architecture

A ResNet-20 model with Bayesian layers was trained on the CIFAR-10 dataset. Bayesian layers added model uncertainty by treating weights as distributions rather than fixed values.

2.2.2 Adversarial Attacks

The Projected Gradient Descent (PGD) attack was implemented using the `torchattacks` library. Adversarial images were generated with parameters: $\epsilon = \frac{8}{255}$, $\alpha = \frac{2}{255}$, and 4 steps. Normalization was applied using CIFAR-10 mean and standard deviation values.

2.3 Experiments

2.3.1 Pre-Adversarial Training

The model was tested on the original and adversarial images. The results showed high accuracy on the original images (89.7%) but a significant drop on adversarial images (0.4%), indicating vulnerability to attacks.

2.3.2 Post-Adversarial Training

The model was trained on adversarial images to improve robustness. After training, accuracy on original images decreased to 80.6%, while accuracy on adversarial images improved to 71.9%, demonstrating enhanced adversarial robustness.

2.3.3 Results of Adversarial Evaluation

The results of the experiments on the CIFAR-10 dataset are summarized below:

- **Accuracy before Training:**
  - Original images: 89.7%
  - Adversarial images: 0.4%
  - Mixed images (original + adversarial): 45%
- **Accuracy after Adversarial Training:**
  - Original images: 80.6%
  - Adversarial images: 71.9%
  - Mixed images (original + adversarial): 76.4%
- **PGD Attack Performance:** The PGD attack was successful in significantly reducing accuracy on adversarial examples with $\epsilon = \frac{8}{255}$, demonstrating the vulnerability of the model.
- **Visualization:** Variance plots were visualized before and after adversarial training to assess changes in uncertainty. Class distribution analysis was also performed to evaluate which classes were more vulnerable to adversarial attacks. These visualizations confirmed that adversarial training led to improved robustness, particularly for classes with higher misclassification rates.

3 Internship Work

- **Company:** Rashmi Metaliks Ltd.
- **Internship Duration:** 8th Jan 2025 – 26th April 2025
- **Location:** Kharagpur, West Bengal
- **Internship Role:** Graduate Engineer Trainee
- **Reporting To:** CEO and President

Introduction

This phase summarizes the work carried out during my internship at Rashmi Metaliks Ltd., where I contributed to multiple strategic projects involving enterprise planning, ERP tracking, automation, and digital reporting. I worked under the Digital Transformation and CEO Cell teams, engaging with top-level management and external vendors.

3.1 Planning and Forecasting – Requirement Alignment

My initial responsibility was to support the evaluation and selection of Planning and Forecasting tools for the organization. Key contributions:

- Built a Proof of Concept (PoC) summarizing:
  - Rashmi's current operational status
  - Core business user requirements
  - Automation gaps in data flow and reporting
  - Comparative overview of vendor offerings.
- Coordinated between external vendor teams (account directors, solution engineers) and internal business users.
- Ensured that proposed solutions aligned with every stated business requirement.
- Organized and facilitated product demos, consolidating stakeholder feedback.

3.2 CEO Cell – ERP MIS and Cost Reduction

Following the PoC, I transitioned to the CEO Cell's cost optimization team, focusing also on ERP MIS and implementation oversight. Responsibilities included:

- Tracked onboarded and in-pipeline statuses for all ERP tools and outsourcing contracts:
  - ERP
  - MES
  - Payroll
  - Planning & Demand Forecasting

  - Contractual Labour Compliance and Management
  - Diesel & Fleet Management
  - Asset Buy/Sell
  - HRMS
  - Transport Management System
- Maintained progress records for both in-pipeline and active ERP solutions along with outsourcing contracts.
- Drafted daily briefs for senior executives outlining:
  - Current status
  - Next action items
  - Target dates
  - Process delays and causes
  - Strategic insights and risk flags
- Consolidated feedback from both internal and vendor teams for management decisions.

3.3 Dashboard Development and Automation

To streamline reporting and remove manual bottlenecks, I led two critical initiatives:

3.3.1 ERP MIS Dashboard Website

- Developed a dynamic dashboard to visually represent:
  - ERP and contract statuses
  - Progress percentage of onboarded tools
  - Financial metrics – project cost, invoiced vs. pending values
  - Collateral tracking (E-ABG, etc.)
- Enabled management to access real-time updates and streamline decision-making.

3.3.2 HR LOI Automation

- Identified inefficiencies in the LOI creation process handled manually by HR.
- Created a Python automation script to:
  - Parse candidate data from Excel
  - Dynamically generate LOI documents
  - Email final LOIs to respective candidates
- Reduced LOI processing time from several days to a few minutes.

3.4 Technologies and Tools Used

- Python (automation scripting)
- Web development stack (used in dashboard creation)
- Google Sheets, Drive, Docs for coordination
- Email automation tools

3.5 Conclusion

The internship at Rashmi Metaliks provided me with extensive exposure to enterprise systems, process alignment, and automation in a corporate environment. It enhanced my technical, coordination, and analytical skills while giving me firsthand experience in handling strategic execution and executive-level reporting.

4 References

1. Y. Feng, T. G. Rudner, N. Tsilivis, and J.Kempe (2024). Attacking bayes: On the adversarial robustness of bayesian neural networks. arXiv preprint arXiv:2404.19640.
2. Pathak, S., Shrestha, S. and AlMahmoud, A., 2024, October. Model Agnostic Defense against Adversarial Patch Attacks on Object Detection in Unmanned Aerial Vehicles. In 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 2586-2593). IEEE.