Dipkamal Bhusal

db1702@rit.edu Google Scholar Github Webpage LinkedIn

Research Overview

I am a Ph.D. candidate and machine learning researcher highly interested in building reliable and trustworthy AI systems. My research focuses on understanding deep learning models through explainable and adversarial machine learning. I have worked extensively in computer vision and cybersecurity, and am currently exploring the reliability and interpretability of large language models (LLMs). My goal is to develop principled tools and techniques that enhance the transparency, robustness, and real-world deployment of modern AI systems..

Research Interests

Explainable AI, Adversarial machine learning, Concept-based explanations, Feature attribution methods, Reliable deep learning, Vision and language models

Experience

August 2021–Present **Graduate Research Assistant**, *Rochester Institute of Technology*, Ai4SecLab, Primary research lies at the intersection of machine learning and security.

- **"Fixing spurious features learned by a model":** I am investigating whether we can utilize post-hoc concept explanations to inspect spurious features learned by image classifiers and guide the model in unlearning them using few-shot examples.
- **"Concept-Driven Adversarial Defense":** I proposed a new defense against adversarial patch attack that combines concept based interpretability to suppress the most influential concept activation vectors during inference. The method successfully neutralizes adversarial patches regardless of size or location, setting a new state-of-the-art on Imagenette. I mentored two students on this project.
- **"Geometry of concepts in LLM:"** I extended my work on faithful concept explanations to LLMs by studying the *geometric relationships* between different concept vectors in order to map the structure of a LLM's "concept space". This can tell what LLMs learn, and either the statistical concept representations align or diverge from human intuition.
- **"Right for Right Reason:"** CNNs frequently "cheat" by exploiting superficial correlations. In this project, I mentored an undergraduate student to develop a scalable framework that leverages vision-language models to automatically generate semantic attention maps using natural language prompts, and train a model for classification and attention alignment to improve generalization, and reduce shortcut reliance. *Current under-review at Neurips workshop*
- **"Faithful concept extraction technique":** Concept explanations provide human comprehensible form of explanations to image classifiers. However, existing methods lack faithfulness to the underlying model. In this work, I proposed a faithfulness aware concept extraction framework. *Currently under-review at Neurips.*
- **"Interpretability price of pruning":** While pruning creates efficient models, its effect on interpretability is unclear. In this project, I mentored a graduate student at RIT to evaluate interpretability of ResNet-18 ImageNette model across pruning levels. We found that moderate pruning improves saliency map faithfulness and sharpens human-aligned attention, while excessive pruning merges concepts and undermines interpretability.
- **"Feature-map smoothing for saliency map interpretability":** In this work, I investigated the trade-off between stability and sparsity of saliency maps and proposed the use of a smoothing layer during adversarial training to obtain high quality saliency maps. *Currently under-review.*

- **"Uncovering feature interactions in image classification":** In this work-in-progress paper, published at NeurIPS'24 ATTRB Workshop, we developed a technique that leverages the Hessian matrix to detect and attribute pairwise feature interactions in image classifiers. This was my first paper as a mentor. We have extended the work with attribution axioms and submitted a full paper. *Currently under-review.*
- **"Benchmarking large language models":** We introduced a knowledge-intensive framework called 'SECURE (Security Extraction, Understanding & Reasoning Evaluation)', a benchmark designed to assess LLMs performance in realistic cybersecurity scenarios. This work was published in ACSAC'24. We extended the knowledge intensive LLM evaluation framework proposed in SECURE and designed LLM benchmarks for CTI-specific tasks, called CTIBench, published at NeurIPS'24. Currently used by Google, Cisco, Trend Micro, and other companies.
- **"Unsupervised adversarial detection":** In this work, I developed a practical method for utilizing sensitivity of model prediction and feature attribution to detect adversarial attack on various deep learning models. Published in EuroS&p 2024.
- **"Modeling explainability":** In my first deep dive into explainable machine learning, I provided a comprehensive analysis of feature attribution based explanation methods and demonstrated their efficacy in different security applications. This work was published as an SoK in ARES'23.

Dec'16-Jun'21 **Co-founder/Software Engineer**, *Paaila Technology*, Co-founder of an AI startup in Kathmandu.

- **Dec'16-Nov'17 (ML engineer):** Contributed as a ML engineer in design and development of robotics and AI solutions. I primarily worked in face recognition and speech synthesis projects.
- **Dec'17-Nov'18 (Project manager):** As the team grew in size, I took the position of project manager to manage the team, and product development.
- **Dec'19-Jun'21 (Managing director):** I took the role of director and was involved in planning and managerial activities, and was responsible for the overall growth of the startup. Due to internal conflicts, I resigned in July 2021 to pursue PhD in the USA.

Sep'20–Aug'21 **Lecturer**, *IIMS College*.

- As a lecturer in the Computer Science department, I taught two BSc. IT undergraduate courses: Introduction to Python and Machine Learning for first and second year students.
- I conducted machine learning workshop for fourth year students.

Education

Ph.D. **Rochester Institute of Technology,** Rochester, NY, USA,
2021–Present *Concentration – Computing and Information Science*
Advisor – Dr. Nidhi Rastogi .

M.Sc. **Tribhuvan University, Institute of Engineering, Pulchowk Campus** Lalitpur, Nepal,
2019–2021 *Concentration – Information and Communication Engineering*
Advisor – Dr. Sanjeeb Prasad Panday
Thesis – Multi-label classification of thoracic diseases using DenseNet on chest radiographs .

B.E. **Tribhuvan University, Institute of Engineering, Pulchowk Campus** Lalitpur, Nepal,
2012–2016 *Electronics and Communication Engineering*
Advisor – Dr. Nanda Bikram Adhikari
Thesis – Prototyping of a voice command based object recognizing robot using speech and image feature extraction .

Publications

[1] A Mehrotra, **Dipkamal Bhusal***, N Rastogi, *"Hessian Sets: Uncovering Feature Interactions in Image Classification"* at Attributing Model Behavior at Scale (ATTRIB), NeurIPS 2024. ***Mentorship**.
[2] MT Alam*, **Dipkamal Bhusal***, L Nguyen, N Rastogi, *"CTIBench: A Benchmark for Evaluating LLMs in Cyber Threat Intelligence"* at NeurIPS 2024 (Spotlight paper. Top 2-3% of accepted papers. *Equal Contribution).
[3] **Dipkamal Bhusal***, MT Alam*, L Nguyen, A Mahara, Z Lightcap, R Frazier, R Fieblinger, GL Torales, N Rastogi *"SECURE: Benchmarking Generative Large Language Models for Cybersecurity Advisory"* at 40th Annual Computer Security Applications Conference (ACSAC). *Equal Contribution.
[4] **Dipkamal Bhusal**, MT Alam, MK Veerabhadran, M Clifford, S Rampazzi, N Rastogi. *"PASA: Attack Agnostic Unsupervised Adversarial Detection using Prediction & Attribution Sensitivity Analysis"* at 9th IEEE European Symposium on Security and Privacy (2024)
[5] **Dipkamal Bhusal**, R Shin, AA Shewale, MK Veerabhadran, M Clifford, S Rampazzi, N Rastogi. *"SoK: Modeling Explainability in Security Analytics for Interpretability, Trustworthiness, and Usability."* at 18th International Conference on Availability, Reliability and Security (2023)
[6] MT Alam, **Dipkamal Bhusal**, Y Park, N Rastogi, *"Looking Beyond IoCs: Automatically Extracting Attack Patterns from CTI"*. 26th International Symposium on Research in Attacks, Intrusions and Defenses (2023)
[7] S Kasarapu, **Dipkamal Bhusal**, N Rastogi, SM Pudukotai Dinakarrao, *"Comprehensive Analysis of Consistency and Robustness of Machine Learning Models in Malware Detection"* Great Lakes Symposium on VLSI 2024.
[8] **Dipkamal Bhusal**, N Rastogi, *"Adversarial Patterns: Building Robust Android Malware Classifiers"*, ACM Computing Surveys 2025

Under Review

[1] **Dipkamal Bhusal**, M Clifford, S Rampazzi and N Rastogi, *"FACE: Faithful Automatic Concept Extraction"*.
[2] A Mehrotra, **Dipkamal Bhusal***, M Clifford, and N Rastogi, *"H-Sets: Uncovering feature interactions in image classifiers"*.***Mentorship.**
[3] **Dipkamal Bhusal**, MK Veerabhadran, M Clifford, S Rampazzi and N Rastogi, *"Towards improving sparsity and stability of saliency maps using feature map smoothing"*.
[4] MT Alam, **Dipkamal Bhusal**, N Rastogi, *"Revisiting Static Feature-Based Android Malware Detection"*.
[5] Sanish Suwal, **Dipkamal Bhusal***, Michael Clifford, N Rastogi, *"Do Sparse Subnetworks Exhibit Cognitively Aligned Attention? Effects of Pruning on Saliency Map Fidelity, Sparsity, and Concept Coherence"*. ***Mentorship.**
[6] Ryan Yang, **Dipkamal Bhusal***, N Rastogi, *"Learning to Look: Cognitive Attention Alignment with Vision-Language Models"*. ***Mentorship.**
[7]Ayushi Mehrotra, Derek Peng, **Dipkamal Bhusal***, N Rastogi, *"Concept-Based Masking: A Patch-Agnostic Defense Against Adversarial Patch Attacks"*. ***Mentorship.**
[8] **Dipkamal Bhusal**, Pradeep Bajracharya, Sanish Suwal, N Rastogi, *"The Evolving Geometry of Concepts: A Layer-wise Analysis of Disentanglement in LLMs"*.

arXiv

[1] **Dipkamal Bhusal**, SP Panday, "Multi-Label Classification of Thoracic Diseases using Dense Convolutional Network on Chest Radiographs"
[2] MT Alam, **Dipkamal Bhusal**, Y Park, N Rastogi, "CyNER: A Python Library for Cybersecurity Named Entity Recognition". on arXiv

Peer-Reviewer

2024 2nd Workshop on Attributing Model Behavior at Scale, NeurIPS 2024.
2024 IEEE Transactions on Artificial Intelligence.
2024 Journal of Artificial Intelligence Research.
2025 Computational linguistics.
2025 Conference on Neural Information Processing Systems (NeurIPS).
2025 Journal of Artificial Intelligence Research.

Teaching and Mentorship

Lecturer Lecturer of Explainable Artificial Intelligence (DSCI 789) at RIT (Fall 2024).

Mentorship Mentored students on research projects.

- *Ayushi Mehrotra, Troy High School:* Research on interactive feature attribution method for image classifiers.
- *Sanish Suwal, RIT Masters in Computer Science:* Supervised his final year capstone project on study of explanation methods on different model training strategies.

- *Sayali Rajesh Kale and Achyut Sridhar Kulkarni, RIT Masters in Data Science:* Supervised their data science capstone project on concept-based explanations.
- *Ryan Yang, Brown University:* Mentored Ryan who visited RIT as a undergrad researcher on XAI research specifically, designing CNN models which are right for right reasons.
- *Ayushi Mehrotra (California Institute of Technology), & Derek Peng (University of California, Berkeley):* Mentored Ayushi and Derek on building interpretability based defense against adversarial patch attack.

Skills

Languages	Python
Frameworks	PyTorch, Keras
Libraries	NumPy, Pandas, Scikit-learn, OpenCV, Matplotlib
Utilities	Jupyter Notebook, Visual Studio, Git, Latex
Electronics	PCB Design, Circuit Simulation, 8 bit microcontroller programming
Industry	Teaching, Project Management, Public Speaking, Business Strategy and Development

Honors and achievements

Scholarship	Financial Support for Ph.D. in Computer Science at RIT, 2021-Present.
Training	Conducted two-week training in Python and Data Science at IIMS College, Kathmandu, 2021.
Award	National ICT Innovation Award by Ministry of Communication and Information Technology (Nepal Government) for Paaila Technology, 2019 .
Scholarship	Full scholarship for Masters in Information Engineering at Pulchowk Campus, Tribhuvan University, 2019.
Award	Most Creative Business of Nepal by Antarprena. Represented Team Nepal at global finals in Copenhagen, Denmark, 2018.
Award	Best Startup of Nepal by ICT Magazine, 2017.
Contest	Winner of Object Oriented Programming Competition by FlipKarma at Pulchowk Campus, 2014.
Award	Secured third position at national level quiz competition Quizmania broadcasted on national television (2012).
Scholarship	Full scholarship for bachelor in engineering at Pulchowk Campus, Tribhuvan University, 2012.
Scholarship	Full Scholarship at Balkumari College, Chitwan (10+2 college equivalent to junior and senior high school level in US) + College Topper + First Position at final examinations in the whole district.

Graduate Courses

Quantitative Foundations, Deep Learning, Statistical Machine Learning, Foundation of Algorithms, Software Engineering, Neural Network, Image Processing, Big Data.

Selected news media

Republica Network: Using AI for better customer experience.

Digital Trend: On waiter robots of Nepal.

AFP News Agency: Nepal's first robot waiter ready for orders.

Kantipur (Nepali): Paaila Technology.

NDTV: In A First, Nepal's Restaurant Uses Robots As Waiters.

India Today: On waiter robots of Nepal.

South Chine Morning Post: Meet Ginger: Nepal's first robot waiter is ready to take your order .

New Business Age Magazine: Embracing the Age of AI, Robotics and ML in Nepal.

Republica Network: Ginger Robot of Nepal.

Kathmandu Post: Emergency Ventilators.

Selected Certifications

1. AI for Medical Treatment, Medical Diagnosis, and Medical Prognosis by deeplearning.ai on Coursera. Certificate earned at June, 2020.
2. Deep Learning Specialization by deeplearning.ai on Coursera. Certificate earned in June 2020. Courses include Neural Networks and Deep Learning, Improving Deep Neural Networks: Hyperparameter tuning, Regularization and Optimization, Structuring Machine Learning Projects, Convolutional Neural Networks and Sequence Models..
3. Project Management Principles and Practices Specialization by University of California, Irvine-The Paul Merage School of Business on Coursera. Certificate earned at May 2020.
4. Mathematics for Machine Learning: Linear Algebra and Multivariate Calculus by Imperial College London on Coursera. Certificate earned at April 2020..

References

Nidhi Rastogi,
Assistant Professor,
Department of Software Engineering, GCCIS, RIT,
nidhi.rastogi@rit.edu.

Michael Clifford,
Principal Researcher,
Toyota InfoTech Labs,
michael.clifford@toyota.com.

Sara Rampazzi,
Assistant Professor,
Department of CISE, University of Florida,
srampazzi@ufl.edu.