Harshitha Machiraju

✉ harshitha.acad@gmail.com harsmac in harshitha-machiraju

About Me

I am an **AI Engineer and Machine Learning Specialist** with expertise in developing and deploying robust **deep learning models** for various applications. My focus includes enhancing model performance against **adversarial attacks**, and **out-of-distribution** scenarios. Recently awarded a **PhD from EPFL**, I have implemented **machine learning** and **AI systems** that address real-world challenges in **computer vision, natural language processing**, and **multi-modal learning**. My experience spans **data preprocessing, model training, parallelization,** and **evaluation** in various testing environments.

Education

Sep. 2019 - Nov. 2024	**PhD in Machine Learning - *EPFL, Switzerland*** Advisors: Prof. Pascal Frossard & Prof. Michael Herzog
Jul. 2014 - Aug. 2018	**B.Tech in Electrical Eng. -** *IIT Hyderabad, India* ***Summa cum Laude*** & Minor in Comp. Sci

Experience

Nov. 2024 - Present	**Independent AI Consultant** (Remote) Designed and trained scalable LLM based systems for health diagnostics and personalized shopping startups.
Sep. 2019 - Nov. 2024	**Doctoral Assistant** *at EPFL, Switzerland* Designed and implemented **robust AI models** for various applications, focusing on efficiency, and resilience to **distribution shifts**.
Sep. 2018 - Aug. 2019	**Research Assistant** *at IIT Hyderabad, India* Developed and deployed **ML models** for autonomous navigation, including the implementation of adversarial testing frameworks.

Projects

- **Adversarial Subspace Analysis in LLMs:** Developed a method to identify **low-dimensional subspaces** within word embeddings that concentrate the most **discriminative** features. Demonstrated that critical information learned by LLMs is often **compactly** represented in these subspaces.
- **Fairness vs. Adversarial Robustness in LLMs:** Demonstrated that **adversarial robustness** d**oes not guarantee fairness**, revealing persistent **biases** in robust LLMs thus highlighting the need for comprehensive fairness evaluations.
- **Lakera's GenAI Security Readiness Report 2024:** Played a key role in the development of the Industry-First AI Security Readiness Report, which provides an **in-depth analysis of organizational preparedness for AI security** in Gen AI applications.
- **Efficient Contrastive Learning for Bias Mitigation:** Proposed CLAD, a novel and efficient contrastive learning-based **training** approach that achieved **State-of-the-Art** on the **Background** challenge dataset. Work published at **BMVC**.
- **Generation of adversarial foggy images for Robustness Evaluation:** Pioneered **GAN-**based creation of adversarial foggy images, marking the forefront of **adversarial weather attack** exploration within this domain. Work published at **WACV**.

- **Enhancing Neural Network Robustness via Latent Perturbations:** Proposed a novel **adversarial training** method based on perturbations in the latent space to increase the robustness of neural networks. Work published at **IJCAI**.
- **Test time Input Processing against Image Corruptions:** Proposed EREN, a novel, **differentiable image processing algorithm** tailored to the **spectral biases of models.** EREN enhances model robustness against **diverse image corruptions** and achieves **superior** performance.
- **Automating Out-of-Distribution Sample Generation by Leveraging Model Biases:** Proposed MUFIA, an innovative algorithm **automating the generation of out-of-distribution samples** by harnessing model **spectral biases**. This work represents a significant advancement in the field, characterized by its utilization of spectral biases for the generation of adversarial image corruptions.
- **Metric design for Robustness Evaluation under varying Weather Conditions:** Pioneered a new metric to gauge the **robustness** of **object detection networks** within navigation systems across diverse weather conditions. **Oral presentation** at **ICIP**.

Selected Publications

- **HM**, M. Herzog, P. Frossard, "Eren: Enhancing deep learning robustness through image pre-processing," (Under Review), 2024.
- **HM**, M. Herzog, P. Frossard, "Frequency-based vulnerability analysis of deep learning models against image corruptions," (Under Review), 2023.
- **HM,** O. Choung, M. Herzog, P. Frossard, "Empirical advocacy of bio-inspired models for robust image recognition," **CVPR** NeuroVision Workshop, 2022.
- K. Wang, **HM**, O. Choung, M. Herzog, P. Frossard, "CLAD: A contrastive learning based approach for background debiasing," **BMVC**, 2022.
- **HM**, V. Balasubramanian, "A Little Fog for a Large Turn," **WACV**, 2020.
- N. Kumari, M. Singh, A. Sinha, **HM**, B. Krishnamurthy, V. Balasubramanian, "Harnessing the Vulnerability of Latent Layers in Adversarially Trained Models," **IJCAI**, 2019.
- **HM**, S. Channappayya, "An Evaluation Metric for Object Detection Algorithms in Autonomous Navigation Systems and its Application to a Real-time Alerting System," **ICIP**, 2018 (Oral).

*Complete List on Google Scholar

Skills

Programming	Python, C, C++, Java, Matlab, SQL, Kubernetes, Docker, Slurm
Frameworks	Pytorch, Tensorflow, LangChain, WandB, Hugging Face, Git, Latex, Illustrator
Languages	English (Native), French (Basic), Korean (Int.), Hindi (Native), Telugu (Native)
Certifications	BlueDot Impact Intro to Transformative AI, AI Alignment, AI governance

Awards and Recognition

- **DeepVision** Grant 2019-2021.
- Qualified for **JICA Scholarship**, 2018.
- **JENESYS Scholarship** 2017, **KVPY** 2013.
- **Special Recognition for a Young Team**, IEEE SP CUP, 2016.
- **Top 10 teams of IEEE SP CUP**, 2016.
- **Academic Excellence Award**, IIT Hyderabad, 2014.

Community Service

- **Reviewer** for ECML, CVPR, TIP, ICVGIP.
- **TA** for Signal Processing & Deep Learning courses.
- **Supervision** of many Masters students projects.