Avinash Amballa

6095054919 | amballaavinash@gmail.com | amballaavinash.github.io | www.linkedin.com/in/avinashamballa

Education

University of Massachusetts Amherst | MS Computer Science | **CGPA:4.0/4.0** Aug 2023 - May 2025
Relevant coursework: Reinforcement Learning, Advanced Natural Language Processing (NLP), Intelligent Visual Computing (3D Deep Learning)

IIT-Hyderabad | B.Tech in Electrical Engineering with minor in Computer Science | **CGPA:8.8/10.0** Jul 2017 - June 2021
Relevant coursework: Data Structures, Algorithms, DBMS, Machine learning, Representation Learning, Linear Algebra

Professional Experience

Google | Student Researcher | Technologies: Python, Pytorch, HuggingFace Feb 2024 – May 2024

- Experimenting arithmetic sampling, to sample diverse sequences in parallel from **Large Language Models** (LLMs) with Chain of Thought self-consistency and MBR decoding strategies on GSM8K and newstest2019 benchmarks with **LLaMa-2**, **Gemma**, and **Flan-T5** models.
- Integrating to **HuggingFace** with **PyTorch.** Incorporating diverse measures of sequence similarity using **BERT** and Box embeddings.

Bosch (AIShield) | Research Scientist | Technologies: Tensorflow, Pytorch, scikit-learn, Docker, Git Aug 2021 – July 2023

- Spearheaded research in **responsible AI** & **AI Security**, focusing on vulnerability assessment, robustness, interpretability, fairness, causality, and drift detection across **ML models** and **DNNs** in computer vision, time series, speech, and natural language processing tasks.
- Developed novel attack and defense strategies for adversarial, poisoning, and model extraction attacks. Published **1 paper** and **4 patents**.
- Played a pivotal role in securing LLMs by focusing on **LLM alignment** and analyzing jailbreaking attacks, developing an application to secure **generative AI models** (AIShield Guardian). Currently used by **5+ organizations**.
- Established **partnerships** with Databricks and Whylabs to enhance AI model security, yielding a **revenue surge of 10%**. Fostered internal partnerships with **2 teams** in the Healthcare sector, to assess vulnerability and improve reliability.
- Built **microservices**, **end-to-end pipelines**, and logging infrastructure across **Azure** & **AWS**, accounting for **30% of the overall workload**.
- Created a **Python library** (PyPI) on adaptive batch size for training AI models. Currently adopted by **15+ researchers**.

GE Digital | Software Development Intern | Technologies: HuggingFace, pandas, Flask, ReactJS May 2020 – July 2020

- Migrated the **web translation** pipeline based on XML and JSON to a fine-tuned **T5 Transformer** on **Tensorflow** and HuggingFace.
- Achieved a **BLEU score of 0.29.** Deployed scalable REST APIs with **Flask**, integrated with **React** interface to demonstrate web translation.

Academic & Research Projects

Aligning LLMs towards safety and helpfulness | UMass Feb 2024 - May 2024

- Fine Tuning LLMs on PKU-SafeRLHF and UltraFeedback datasets with DPO and **Q-LORA** to align toward safety and helpfulness.

Optimization in Reinforcement Learning | UMass Sep 2023 - Nov 2023

- Programmed Reinforce with baseline, Actor-Critic, Episodic Semi Gradient n-step SARSA in **PyTorch** for Acrobat, Cartpole environments.
- Attained stabilized **mean rewards of 470 (max:500), -100 (max:0)** on Cartpole and Acrobat respectively using Reinforce and Actor-Critic.

Gyro Correction in IMU sensors | IIT-Hyderabad, DRDO India Apr 2021 - Jul 2021

- Built a gyro correction model for **IMU sensors** to mitigate noise and axis misalignment, employing diverse architectures such as DB LSTM, **LSTM** with attention, and **Transformer** Encoder. Trained on EUROC dataset with Huber Loss.
- Achieved **validation loss of 0.229** with attention models surpassing SOTA Dilated CNN's validation loss of 0.246 with **hyperparameter tuning**.

Explaining Adversarial Robustness | IIT-Hyderabad Jan 2021 - Apr 2021

- Analyzed the learned Convolution filters and visual explanations (**SHAP, CAM)** pre and post-adversarial training across **AlexNet** and **ResNet**.
- Examined **Fourier analysis** on adversarial examples across **3 datasets**. Found no correlation between frequency and adversarial behavior.

ViCaP: VIdeo Captioning And Prediction | IIT-Hyderabad Sep 2020 - Dec 2020

- Implemented a video captioning method, utilizing a pre-trained **VGG16** feature extraction with attention based **encoder-decoder LSTM** model.
- Trained on MSVD dataset with cross-entropy loss. Achieved a higher **BLEU-4 score of 0.67** compared to a baseline with CNN and LSTM.
- Predicted the missing video frames through **pix2pix conditional GAN**. Investigating self-supervised learning techniques for the same.

AlphaConnect-4 | IIT-Hyderabad Jan 2020 - Apr 2020

- Created competitive **multi-agent Reinforcement Learning** on connect-4 game, utilizing **MCTS** for opponent and **Actor-Critic** for agent.
- Designed the game environment in **Python**. **Fine-tuned** the learned connect-4 agent on the connect-5 game to improve its performance. .

Technical Skills - Machine learning / Data Science

Programming Languages: Python, C, C++, JavaScript, HTML | Familiar: Java, R, SQL, CSS

Tools/Libraries: PyTorch, TensorFlow, Keras, Scikit Learn, Numpy, Pandas, Matplotlib, Scipy, OpenCV, OpenAI gym, NLTK

Software/Frameworks: Git, Docker, Flask, Node.js, jQuery | Familiar: Azure, AWS, React, Elasticsearch, PostgreSQL, DevOps

Publications & Preprints

[1] Govindarajulu, Y., **Amballa, A.,** Kulkarni, P., & Parmar, M. (2023). Targeted Attacks on Time Series Forecasting. arXiv:2301.11544.

[2] Amballa, A., Sasmal, P., & Channappayya, S. (2022). Discrete Control in Real-World Driving Environments using Deep Reinforcement Learning. arXiv:2211.15920.

[3] Amballa, A., Mekala, A., Akkinapalli, G., Madine, M., Yarrabolu, N. P. P., & Grabowicz, P. A. (2024). Automated Model Selection for Tabular Data. arXiv:2401.00961.

Patents

[1] IN Patent # 202241068482: “A method to detect poisoning of an AI Model and a System thereof.”

[2] IN Patent # 202241065028: “A method of Targeted Attack on Time Series Models to alter the DIRECTION”

[3] IN Patent # 202241065034: “A method of Targeted Attack on Time Series Models to alter the MAGNITUDE”

[4] IN Patent #202441006640**:** “A method of Sponge attack on Deep Learning Models to increase the inference time”