Avinash Amballa

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

Education

University of Massachusetts Amherst | MS Computer Science | **CGPA:4.0/4.0** Aug 2023 - Dec 2024
Relevant coursework: Advanced Natural Language Processing (NLP), Intelligent Visual Computing, Reinforcement Learning, Statistics

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, Image Processing

Professional Experience

Dedrone, USA | ML Engineer Internship | | Technologies: Python, Pytorch, Flax, CUDA, scikit-learn June 2024 - Aug 2024
- Implementing track recognition from time series RF and radar data by employing architectures such as **1D CNN, LSTM, GCN, Transformers (Sparse, linear attention), State Space Models (S4, Mamba), xLSTM** to handle the long range dependence in linear time.

Google, USA | Graduate Student Researcher | Technologies: LLM, Python, Pytorch, HuggingFace Feb 2024 – May 2024
- Experimented with arithmetic sampling to generate diverse sequences in parallel from **Large Language Models** (LLMs) with Chain of Thought self-consistency (**LLaMa-2**, **Gemma** on GSM8K, StrategyQA) and MBR decoding (**Flan-T5, MT0** on WMT14) strategies.
- Integrated to **HuggingFace Transformers** in **PyTorch.** Achieved a **3-5%** improvement in accuracy with CoT self consistency on the GSM8K.

Bosch (AIShield), India | Research Scientist | Technologies: Responsible AI, Tensorflow, Pytorch, scikit-learn, Docker Aug 2021 – July 2023
- **Published paper and 4 patents** as a result of research on vulnerability assessment (robustness), interpretability, fairness, causality, and drift in **ML models** and **DNNs** across **7+ tasks** from computer vision, time series, speech, and natural language processing.
- Innovated attack and defense strategies for **adversarial**, **membership inference**, poisoning, and model extraction attacks.
- Secured LLMs by focusing on **LLM alignment** and analyzing jailbreaking attacks, developing AIShield Guardian application **used by 5+ organizations** to secure generative AI models
- Established **partnerships** with Databricks, Whylabs and **2 teams** in healthcare to enhance AI model security, yielding a **revenue surge of 10%**
- Designed **microservices**, **end-to-end pipelines**, 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 which was adopted by **15+ researchers**.

GE Digital, India | Software Development Intern | Technologies: TensorFlow, HuggingFace, Pandas, Flask, ReactJS May 2020 – July 2020
- Migrated **web translation** pipeline based on XML and JSON to a fine-tuned **T5 Transformer** on **Tensorflow** and HuggingFace.
- Achieved a **BLEU-4 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
- Aligned LLMs (**LLAMA-2 7B**) toward safety using PEFT (**LORA, Prompt Tuning**) on PKU-SafeRLHF benchmark with **SFT, RAFT, DPO**.
- Scored **93%** safe on DPO model with Llama-Guard vs. SFT's **40%** on I-CoNa. DPO achieved **63.3%** performance vs. SFT's **60.38%** on PIKA.

Motion Synthesis in Latent Space | UMass Feb 2024 - May 2024
- Generated text-to-motion sequences in latent space using **GANS, Diffusion** coupled with **VAE** and **CLIP** on the HumanML3D benchmark.
- Demonstrated that simple GAN architecture with three linear layers in the latent space achieves **FID of 2.39** and **diversity score of 8.92**

Python Question Answering with Gemma | Kaggle Feb 2024 - Apr 2024
- Fine-tuned **Gemma** on StackOverflow Python questions and coupled with **RAG** framework on vector database with **CoT** prompting.

Optimization in Reinforcement Learning | UMass Sep 2023 - Nov 2023
- Programmed Reinforce with baseline, Actor-Critic, Semi Gradient n-step SARSA, Evaluation strategies (BBO) in **PyTorch** for Acrobat, Cartpole.
- Attained stabilized **mean rewards of 470 (max = 500), -100 (max = 0)** on Cartpole and Acrobat respectively using Reinforce and Actor-Critic.

Gyro Correction in Inertial measurement unit (IMU) sensors | IIT-Hyderabad, DRDO India Apr 2021 - Jul 2021
- Built a gyro correction model for **IMU sensors**, employing architectures such as DB LSTM and **BERT** Encoder.
- Trained on EUROC data with Huber Loss, attaining **validation loss of 0.229** with BERT surpassing SOTA Dilated CNN's val loss of 0.246.

ViCaP: VIdeo Captioning And Prediction | IIT-Hyderabad Sep 2020 - Dec 2020
- Implemented video captioning method, utilizing a pre-trained **VGG16** with attention-based **encoder-decoder LSTM** model on MSVD dataset.
- Trained with cross-entropy loss to attain **BLEU-4 score of 0.67** and predicted missing video frames through **pix2pix conditional GAN**.

Publications, Preprints & Patents (Google scholar: scholar.google.com/citations?user=wi6Fpr4AAAAJ&hl=en)

[1] Arithmetic sampling with CoT self consistency and MBR decoding under review
- Attained a 3-5% point increase in accuracy on the GSM8K dataset and a 1-6% point increment in BLEU score for WMT14 tasks.

[2] Targeted Attacks on Time Series Forecasting arXiv:2301.11544, IN Patent App No. 202241065028, 202241065034
- Introduced Directional, Amplitudinal, and Temporal targeted adversarial attacks on time series forecasting models.

[3] Discrete Control in Real-World Driving Environments using Deep Reinforcement Learning. arXiv:2211.1592
- Trained Self-Driving cars in multi-agent RL framework, which effectively transfers real-world driving environments into gaming simulations.

[4] Automated Model Selection for Tabular Data arXiv:2401.00961
- Developed an framework that incorporates feature interactions using Priority-based Random Grid Search and Greedy Search methods

[5] A Method to detect AI poisoning attacks; A Method of Sponge attack IN Patent App No 202241068482, 202441006640

Ongoing: Data Free Model Stealing, Diffusion Prior for Anomaly Detection, Superposed decoding in NLI, Physics Informed Neural Networks

Technical Skills - Machine learning / Data Science / ML System Design

Programming Languages: Python, C, C++ | Familiar: CUDA, Java, R, SQL, JavaScript, HTML, 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, System Design, PostgreSQL, DevOps