BASU VERMA

(+91)-8582975408 | (+91)-7250074110
basuverma120299@gmail.com
linkedin.com/in/basu-verma-219558312/
github.com/vermabasu?tab=repositories
vermabasu.github.io/profile

SUMMARY

Experienced AI Engineer with expertise in CV, RL and DL, Image Processing, Natural Language Processing and GenAI. With a proven track record of developing and deploying cutting-edge AI solutions, I have successfully tackled complex challenges across diverse domains. My passion is to take advantage of advanced technologies to drive innovation and I am dedicated to delivering impactful results while collaborating seamlessly within teams to achieve shared goals.

EDUCATION

MTech, Data Science	Indian Institute of Technology(IIT)	May 2022
BTech, Electronics & Instrumentation	West Bengal University of Technology, W.B(IN)	May 2020

TECHNICAL SKILLS

MLOps / LLMOps:	MLFlow, Github Actions, DvC, Iterative Studio, cron job, scheduler, YAML scripting.
Azure MLOps:	Databricks, AzureML, Terraform, AzureDevOps, ARM, RBAC, Secure CI/CD practices.
Cloud:	Azure, GCP, Docker, Kubernetes, Edge AI.
GenAI:	RAG, LangChain, LLamaIndex, DeepSpeed, Gaurdrail, VectorDatabase, LLM's, LoRA, Bedrock
CV:	OpenCV, Image Processing, Object Detection, Segmentation, Tracking, YoLo Family
Niche:	Reinforcement Learning (Q-learning, DQN, Double DQN, TD3, PPO, TRPO, A2C, DDPG)
Languages:	SQL, Python (pandas, numpy, seaborn), HTML, CSS, JS
Other Tools:	Flask, Streamlit, Gradio, Excel, Power Point, AWS (EKS, EC2, S3,Sagemaker)

PUBLICATIONS

- Verma, Basu, et al. "SIGN-Diffusion: Generating User Specific Online Signature for Digital Verification." International Conference on Pattern Recognition. Cham: Springer Nature Switzerland, 2024.
- Verma, Basu, et al. "SIGN-GAIL: Rewarding Online Signature Generation for Digital Imitation." Proceedings of the Winter Conference on Applications of Computer Vision. 2025.

EXPERIENCE

HCLTech, Banglore, IN: Technical Lead — Aug 2022 – Present

MLOps/LLMOps

- Developed end-to-end LLMOps/MLOps pipelines with GitHub Actions, MLflow for CI, cron jobs and schedulers for CT, Docker, Kubernetes and SageMaker endpoints for CD, leveraging AWS services.
- Managed Kubernetes based ML deployments, optimizing auto-scaling policies ML and LLM models.
- Implemented CM pipeline with Flask, featuring drift detection visualization, model status tracking, and comprehensive metrics (e.g., token efficiency, latency, BLU & BERT score, Pass@K).
- Implemented model drift and data drift to identify performance degradation and trigger automated retraining pipeline using AWS EC2, S3, Lambda and sagemaker pipelines.
- Fine-tuned multiple large language models (e.g., LLaMA, GPT-Neo) with advanced optimizers like LoRA, QLoRA and deepspeed on distributed environment, achieving up to 40% improvement in model performance.
- Emplopyed distributed architecture utilizing multiple GPU devices with data parallelism and model parallelism.
- Implemented various computer vision, natural language processing, and generative AI use cases to demonstrate the capabilities and performance of operational pipelines.

MLOps pipeline in Lanedetection

- Implemented Convolutional Neural Networks (CNNs) on each frame of the video to extract spatial features and detect lane markings effectively and YOLO (You Only Look Once) architecture for object detection optimizing real-time lane detection.
- Processed large-scale video datasets, utilizing **Kafka Streaming** to efficiently stream high-volume video data from producer to consumer for real-time analysis.
- Developed end-to-end MLops pipeline with github and Azure platform, featuring model and artifacts versioning, retraining, evaluation, continuous monitoring and cloud deployment with docker and kubernetes.
- Employed various responsible AI tenets such as deepcheck, Exlpainable AI using SHAP for local & global explanation.

Multi-modal Brain Tumor Segmentation

- Performed mult-modal BraTS Brain MRI Scans analysis and developed a novel multi-agent reinforcement learning architecture for segmenting the MRI scans for various tumor components.
- Proposed a U-Net like architecture as actor of RL agent, processing MR images and initial segmentation mask, generating updated updated mask ensuring accuracy of segmentation.
- Developed the entire framework in co-operative fashion, with each agent having shared reward resulted in robust and highly accurate model.

PhotoGraphy Recommendation System

- Designed and implemented a Reinforcement Learning (RL) agent capable of recommending optimal camera soft parameter values based on input images, while adapting to individual user preferences to ensure personalized experiences.
- Developed a TD3 RL agent with a custom environment, actor, and critic with RLHF technique, enabling effective training across multiple users and diverse image genres.
- Achieved approximately 15% improved performance over OpenAI models by evaluating the RL agent's results in comparison, highlighting its superior ability to personalize camera parameters.

SigDiff - Signature Generation and verification

- Designed and implemented a novel Sign-Diffusion framework for robust online signature generation and verification, enhancing security for digital transactions.
- Utilized conditional diffusion and state space models to extract intricate spatial and temporal features from signatures, improving resilience against manual and digital forgeries.
- Integrated capabilities to identify system-generated deepfakes, contributing to the advancement of secure online signature verification methods.

ACADEMIC PROJECTS

Driver Drowsiness Detection

April 2021 - May 2022

- Developed drowsiness detection combining Haar-Cascade and CNN for blink frequency classification, with YoLo-v5 for eye detection and blink frequency analysis.
- Incorporated a custom dataset for YoLo model containing 1336 training images with balanced open_eye and closed_eye classes, alongside 414 validation images.
- Integrated YoLov5m and YoLov5l models achieving outstanding accuracies of 98.2% and 98.6%, respectively.
- Demonstrated efficient frame processing times averaging 0.008 seconds using Tesla K80 GPU and 0.53 seconds on Raspberry Pi board 4.