AYUSH VERMA

New Delhi, India (Open to Relocate)

• ayushvrma08@gmail.com • github.com/ayushvrma • linkedin.com/in/ayushvrma • (+91-8604091203)

AI Engineer

CAREER SUMMARY

Engineer with deep experience in building revenue driven production-ready AI systems for computer vision, generative modeling, and LLM agent applications. Proven success in deploying CUDA-optimized inference systems, building scalable pipelines for image/video generation, and integrating agentic intelligence across product stacks. Skilled in end-to-end ownership from research to robust, autoscaled cloud deployments. Immediate Joiner.

PROFESSIONAL EXPERIENCE

Spyne.ai

Gurugram, India (January 2024 - Present)

B2B SaaS with $16M series-A funding transforming image cataloging across the Globe.

AI Researcher, Grade II

Reporting to the Vice President, 5 day in-office working closely with product, tech and data teams on Computer Vision solutions with complete ownership of projects to solve cataloging problems by generating research statements, experimentation, model tracing, dockerisation and scalable CPU-GPU deployments on aws using Target Groups and ECR images serving 450 clients worldwide churning 20M images/month.

Generative:

- **Novel View Synthesis**: Generated upto **72 novel views of object using just 8 images** with by writing a pull based python backend leveraging a SQL database utilising segmentation masks, NVIDIA Triton Python Backend for Colmap (CPU) and VGGSfM (GPU) and custom losses for **camera position estimation** and refinement for text correction taking Spyne's 360 product (view here) to launch. Currently processing 500 SKUs/day for 150 clients worldwide.
- **3D Scene Generation:** Wrote a pull based python pipeline to generate a 3D model of an object using a 360 video utilising initial colmap pointcloud refined using 3D Gaussian Splatting (3DGS), with more experimentations using Deformable Beta Splatting with Diffex Refinement for further artifact refinement, leading to Spyne's 3D upcoming product launch. Also wrote a custom viewer in Visor to restrict camera FOV for 3D scene viewing.
- **Trained, traced and deployed to scale Image2Image:**
  - U2Net (2017) on a dataset generated by a custom trained LORA StableDiffusion Text2Image pipeline written for dataset generation, to add shadow effect on an object.
  - Pix2PixHD (2018) network with changed ResNet layers for better loss on custom trained LORA SDXL Image2Image dataset, to add reflection of the object on the floor it's on.
- **3D Human Talking Head Synthesis:** Deployed an experimental feature for car dealerships by writing a pipeline for multilingual few shot talking head video generation using image(s) or videos and text mapping mel-spectrograms, phoneme sequences, or other acoustic features to 3D talking face mesh. (MultiTalk)
- **Deployed Voice Cloning** to 0.33-second/inference (512 tokens BPE) optimised for a g4dn.2xlarge ASG, parallely processing custom **AI avatars** from photo(s)/ videos adding selectable customisation to Video Product to dealerships. (Bark, OpenVoice)
- **Optimised SDXL inference** to **0.25sec for 50-step** production level **image inpainting** solving removal of unwanted objects on the ground during photoshoots using frameworks like onediff and oneflow on Nvidia Triton Model Repository.
- **Object Detection and Tracking** using yolov7 for USP feature generation, additionally shipped with the Video Product to 40 dealerships.
- **Finetuned and Traced lama** for image outpainting task integrating it in the main image processing pipeline.

Agentic:

Made and finetuned a swarm of **5 LLM-Agents with access to tools** implemented in Langgraph, providing key business insights querying from unifying PostgreSQL, MongoDB, Knowledge Graphs from unstructured conversation data to query data, plot graphs, draw predictions and general conversation on the fetched data backend locally hosted for the organisation.

Engineering:

- **Integrated newrelic** for better error logging in all product (Images, Videos, 360) pipelines for both production, staging and development environments.
- **Trained a Regression Network** for delta y shift for video stabilisation using SuperGlue point tracking.
- **Deployed a View Angle Prediction** network using a CLIP-ViT Backend trained on a blender dataset made by a custom data generation pipeline, replacing the current model used at Spyne with a 4% increase in prediction accuracy and 20% more type coverage quantised to int8 onnx.
- Implemented and integrated an automated novel way to backprocess data and trigger reprocessing cutting custom labor hours by 120/week.
- **MLOps**: Implemented kafka producer-consumer backed redis queue to separate various processing servers for load separation for validation and image processing for CPU servers and autoscalable EKS model pods for triton model deployment.

Diool
Douala, Cameroon (March 2022- September 2022)
Providing financing options to B2B vendors.

AI/ML Intern

Reporting to the CTO, interning remotely as a part of the AI team during my Bachelor's study to analyse financial data, make data driven insights and provide accurate modeling solutions.

- Conducted analysis of the African financial market to provide customizable Buy Now, Pay Later (BNPL) payback schemes for customers.
- Worked with Trumania simulations to create market simulations for research purposes.

RESEARCH EXPERIENCE

University of South Carolina, Columbia, SC (June 2023 – Dec 2023)
NLP Research Intern
Developed a novel pipeline remotely researching under Dr. Usha Lokhala (Google Scholar) for Aphasia Disease (wiki) identification.

- Used BERT and ConceptNet Numbatch to store and classify human responses getting a validation accuracy of 73%.
- Creating ontology aware Knowledge Graphs for better embedding storage.
- Integrated prediction losses in the LLM pipeline for better judgment and classification.

University of Surrey, Guildford, England (May 2022 – June 2022)
Deep Learning Intern
Worked on Fine Grained Sketch-Based Retrieval (FGSBR) remotely researching under Dr. Ayan Kumar Bhunia (Google Scholar) in e-commerce.

- Integrated Pair Triplet Losses in the baseline Siamese network with Spatial Attention to improve accuracy from 88.3% to 94.4% on real world examples.

EDUCATION

Bachelor of Engineering in Computer Science and Business Systems (2020-2024)
Thapar Institute of Engineering and Technology, India
Relevant Coursework:
Linear Algebra, Stats, Machine Learning, Theory of Computation (Gold Medalist), Compiler Design, Systems Design, Software Engineering, Cybersecurity, Design Thinking, Behavioral Economics

Skills

python, Bash, SQL ,PyTorch, FastAPI, CoreML, Triton, OpenCV, SDXL, LLMs, OneFlow, ConceptNet, BERT, Triplet Loss, AWS, Docker, Redis, NewRelic, W&B, ASG ,Git, Blender, Plotly, langraph, langchain, n8n, knowledge graphs, vectorDB, fastapi, langraph, autogen