Ashish Kumar Singh

College Park, MD - 20740 • (240) 886 9207 • ashishkmr472@gmail.com • aksingh4@umd.edu
github.com/AshishKumar4 • linkedin.com/in/aksnip • ashishkumarsingh.com

Machine Learning researcher with 4+ years of experience in diffusion models, NLP, and large-scale distributed systems. Seeking a research/engineering role to develop cutting-edge solutions in computer vision, NLP, and autonomy

EDUCATION

Masters of Science in Applied Machine Learning, **University of Maryland, College Park** (4.0 GPA) — Expected May 2026

Bachelor of Technology in Computer Science, **Vellore Institute of Technology** (8.2/10 CGPA) — Jun 2016 - Jun 2020

TECHNICAL SKILLS

ML Frameworks: JAX, Flax, TensorFlow, PyTorch, PyTorch-lightning, HuggingFace, Pandas, Keras, OpenAI Gym

ML Techniques: Quantization, Pruning, DDP, Distributed Training, Diffusion Models, GANs, Transformers, Generative AI, NLP, Computer Vision, Multi-modal models, Distillation, LoRA, CNNs, Object Recognition, Face Recognition

Programming: Python, C++, Golang, CUDA, C, x86 ASM, JavaScript, TypeScript, OpenGL, WebGL, Bash

Cloud & ML Ops: Google TPUs, GCP/AWS, Weight and Biases (wandb), SLURM, Google BigQuery, Hadoop

WORK EXPERIENCE

Dyte Pvt Ltd, Bengaluru - Machine Learning and Systems Engineer — Jun 2021 - Jun 2024

- **Designed and engineered** advanced **RAG** system with **3x** more accuracy over naive RAG, leveraging **Multi-agent LLMs**, **chain of thoughts**, and **self-auditing** mechanisms
- **Led** development on transcription engine, improving word error rate by **30%** over OpenAI Whisper on noisy meeting recordings
- **Spearheaded** architecture design and development of **WebRTC SFU**/Networking Stack, increasing load handling capacity/scalability by **15x**
- **Engineered voice-to-voice bot SDK** (Deepgram + LLaMA) with **<800ms latency** using speculative execution
- **Developed** LLM powered automations to monitor GitHub repository changes and **auto-generate** reports, cutting manual reporting by **15 hours weekly** and improving code review efficiency by **20%**

Hyperverge, Bengaluru - Machine Learning Researcher — Dec 2019 - Jun 2021

- **Spearheaded** R&D on **state-of-the-art** facial anti-spoofing **CV** models, achieving **ISO 30107-3** certification
- **Designed** rotational invariant **face detection** models using **feature pyramid networks**, improving detection by **40x**
- **Built parallel data processing** and **TPU training pipelines**, reducing training times from weeks to hours**,** achieving a **30x** performance increase
- **Improved facial recognition** precision by **55%** using **contrastive losses** and optimized data augmentations
- **Created** training code for **Progressive Calibration Networks (PCN)** from scratch in **C++** and Caffe to boost production pipeline performance by **25%**

PROJECTS

Research & Open Source

Diff2Lip 2 - Ongoing @ UMD (to be published in ICCV) 🔗 — Oct 2024 - Present

- **Researching** audio-guided **lip-synchronization** with **diffusion** to generate high-fidelity frames
- **Conducted** training and ablation studies on University **HPC SLURM** Clusters, boosting experimentation rates by **32x**
- **Reimplemented** original (diff2lip1) **PyTorch** codebase to **Pytorch lightning**, leveraging **DDP training** techniques to scale training times by **8x**

FlaxDiff - Diffusion Library🔗 — Jun 2024 - Sep 2024

- **Implemented Flax/Jax**-based **diffusion** library with **17+** diffusion techniques akin to **Huggingface Diffusers**
- **Trained 100M-**parameter models on **250M+ images** using **128 TPUv4s** in **DDP** from **scratch**
- **Authored 3** open-source **tutorials** demystifying **diffusion models** and **generative AI** techniques

Facial Anti-Spoofing and Liveness Research @ Hyperverge 🔗 — Apr 2020 - Jun 2021

- **Pioneered** DL techniques for **face anti-spoofing**, pivotal for the company in achieving **ISO 30107-3** certification
- **Innovated** feature extraction backbones using **Feature Pyramid Networks (FPNs)** and **parallel prediction stages** trained on different contrastive losses, increasing model **PR-AUC** by **30%**
- **Introduced** pipeline optimizations via **distributed training** and **Bayesian hyperparameter tuning,** enabling **250+** weekly experiments, boosting model performance by **5x**

ML & AI Projects

CrawlMind - AI driven web crawling engine and OSINT research tool 🔗

OpenAI APIs, Playwright, Beautiful Soup, AsyncIO

- **Accelerated** the web crawling process by **70%** through LLM-guided link prioritization, reducing redundant page visits
- **Structures** scraped data using LLMs to enable **comprehensive analysis and actionable insights**
- **Integrated** multi-source **OSINT** gathering to broaden data coverage and enhance intelligence outcomes.

Cogito - Advanced Multi-agent RAG Assistant System (WIP) 🔗

Python, OpenAI APIs, Gemini APIs, Solr, GraphRAG, Chain-of-Thought Reasoning

- **Implemented** a multi-stage Retrieval-Augmented Generation system that breaks down complex questions into smaller units, improving accuracy in early tests by **~30%** over naive RAG methods
- **Leverages** LLMs and **chain-of-thought** methods to generate roadmap and self-audit results, resulting in highly consistent, accurate results
- **Breaks down** problem into subproblems and builds a **query graph**, predicting node leaves based on roadmaps and current thoughts, which inturn are solved by **multiple LLM agents** in parallel

NeuralGPU - A CUDA DNN library from scratch 🔗

C++, CUDA (handwritten kernels), Custom Keras-like Functional API

- **Developed** a fully functional neural network library with a Keras-like API in **C++/CUDA**
- **Achieved** up to **20× faster training** throughput compared to a naive CPU-based approach in synthetic benchmarks
- **Implements matrix ops**, **auto-differentiation**, and **multi-layer** networks optimized for GPU parallelism using hand written **CUDA** kernels

Other Notable Projects

Aqeous Operating System and Kernel 🔗

X86 ASM, C, AHCI, VESA, SSE, AVX, Multiprocessing

- **Built** a kernel and a simple user-level **OS completely from scratch**, with multiprocessor, multithreading **SMP** and SSE based double buffered composting windowing system using **VESA** drivers for **GUI**, simple shell, **ELF** file support
- **Runs** a ported **FASM** assembler for development and a user level C library for developing software for it, with a simple syscall interface
- **Listed** in the official OS dev forum wiki.

CERTIFICATIONS

Certificate in Artificial Intelligence and Big Data, **National University of Singapore (A+ Grade)** Jun 2019

Certificate in Big Data and Hadoop System Administration, **Hewlett Packard Enterprise** Jun 2019

AWARDS AND RECOGNITIONS

- **1st Place Winner,** Hardware CTF Competition, NullCon Goa, Goa Sep 2022
- **2nd Place Winner,** Hardware CTF, NullCon Goa, Goa Mar 2020
- Invited as **TOP 100 Startups of the Year, StartupIstanbul, Istanbul**, Turkey Oct 2019
- **1st Place Winner, Hackgrid 2019 Hackathon,** ADG VIT Chapter, VIT Mar 2019
- **1st Place Winner, Bio-Inspired Design Challenge,** SBST School, VIT Apr 2017