Dr. Avinash Kumar Singh

Hyderabad, India | +91-9005722861 | avinashkumarsingh1986@gmail.com | http://avinashkumarsingh.in

Profile

With over 14 years in AI, I have evolved through roles as an ML Researcher, Engineer, Product Manager, and now as Chief AI Scientist. I have led the development and deployment of deep learning-based computer vision and NLP models on platforms like AWS, GCP, Humanoid Robots, Edge Devices like Jetson Nano, Raspberry Pi, and NXP boards. My expertise extends to tackling challenges in concurrency, security, and latency. My academic journey, enriched by a Ph.D. and postdoctoral research, provides a profound understanding of neural networks across diverse data types, while my industrial experience ensures practical AI solutions are deployed effectively, serving real users. This unique blend of research and industry expertise enables me to lead in crafting and delivering impactful AI innovations, driving business transformation and societal advancement.

Experience

AI CONSULTANT & CORPORATE TRAINER | ROBAITA, HYDERABAD, INDIA | SEP 24 – CONT...

- Mentored and empowered 1,500+ students and working professionals from diverse backgrounds. Led immersive sessions on Fine-tuning Large Language Models (LLMs), designing robust Retrieval-Augmented Generation (RAG) systems, building custom AI chatbots, defending LLMs against prompt injection attacks, implementing Model Context Protocol (MCP) for efficient context management, and developing Agentic AI solutions.
- Engineered and deployed a cutting-edge visual language model-based computer vision system to drastically reduce loose picking of Apple iPhones in a warehouse environment, achieving a significant reduction from 20% to 5%. This innovative solution leverages visual question answering to enhance product identification, ensuring operational efficiency and accuracy.
- Designed and developed an advanced conversational AI system tailored to resolve complex challenges within the supply chain. This intelligent system facilitates seamless communication with each supply chain node and accurately predicts arrival times by analyzing routes and external factors such as weather disruptions, protests, and regulatory conditions, ensuring timely and compliant operations.
- Conceptualized and implemented "Talk to Your Document," a generative AI solution that enables interactive question-and-answer sessions with historical documents and databases. This solution achieved impressive BLEU and ROUGE scores of 0.85, demonstrating exceptional accuracy in retrieving and synthesizing relevant information.

GLOBAL SOLUTION LEADER | BRANE ENTERPRISES LLP, HYDERABAD, INDIA | MAY 20 – SEP 24

- A LLM model is finetuned on financial reports to understand complex financial reasoning (mostly tables). The system could generate the mathematical formula to process the query, retrieve the argument from the table and compute the value as the output.
- Implemented a Face Recognition-based office attendance system, replacing the existing RFID system, and achieving organization-wide deployment.
- The system serves 2,856 employees with a 97.63% accuracy rate, resulting in annual savings in operational costs.
- Led a groundbreaking project to design smart glasses for visually impaired individuals, providing comprehensive assistance in reading, navigation, currency identification, person recognition, and scene understanding.
- The system can detect obstacle up to 5 feet, can help in reading English and six Indian languages, could recognize 9,605 objects and labels with 89.76% accuracy.
- Successfully delivered a Driver Monitoring System (DMS), incorporating real-time monitoring and safety features. The system tracks driver drowsiness, smoking, drinking, eating, phone usage, and seatbelt compliance, resulting in a 40% reduction in driving violations.
- Captured and recorded over 1,000 violations with images, date, time, and other details.
- Provided live streams from both interior and exterior dash cameras, enhancing monitoring accuracy.

SENIOR RESEARCHER | MONTPELLIER UNIVERSITY, FRANCE | NOV 20- NOV 21

- I was associated with the robotics lab (LIRMM) and worked on the EU Project SOPHIA. As the in-charge of Work package 5, I helped my team to coordinate between different project partners Italian Institute of Technology, Italy, INAIL, Italy, and LIRMM, France for data acquisition, human-robot interaction and to derive a deep learning model for action recognition.
- Designed and developed a sensor agnostic, Bidirectional LSTM based deep neural network for action recognition. The model is tested in the presence of Xsens suit (used for motion capture) and Intel RealSense and Microsoft Kinect RGB-D data (3D skeleton). The research is published in 21st International Conference on Humanoid Robots.
- The model is integrated with KUKA robot to help human in physical assistance, e.g. carrying object, release object, place object etc. in industrial environment by understanding the human actions.

POST DOCTORAL RESEARCHER | UMEÅ UNIVERSITY, SWEDEN | FEB 18 - JAN 20

- During this postdoc, I closely worked with Professor Kai-Florian Richter and Professor Thomas Hellström. I was a part of intelligent robotics lab, during this postdoc, we designed a dialogue based human robot interaction system that allows humans, to talk to the robot. This work was published in a 'A' rated conference ECAI-2020.
- We developed and implemented a robot collaboration framework that enables robots to have dialogues by translating their actions into the natural language. This work was published in Journal of behavioural robotics and also featured in Softbank robotics under the best 20 projects in 2020.

DEPUTY MANAGER | HCL TECHNOLOGIES, NOIDA, INDIA | FEB 17 - JAN 18

- I joined the HCL machine learning division (Noida) when this was a 3 members team. In the span of one year, we conducted 4 successful POCs and grew the team to a 16 members team.
- As a deputy manager my responsibilities were to handle the client interaction, project scoping, find the place where machine learning solutions can be pitched (to integrated with existing workflows) within and outside the organization.
- Helped my team to set up the machine learning GPU infrastructure, sketching project roadmaps, resource allocation and tracking. Further motivating team to follow the software engineering best practices such as maintaining git, hygiene of code (following the coding standards), etc.

ML ENGINEER | ECLERX SERVICES LIMITED, MUMBAI, INDIA | NOV 15 - FEB 17

- eClerx is an Indian IT consulting and outsourcing multinational company. I worked there as a full stack developer and my job was to design and develop NLP, ML solutions. We used image pre-processing to improve OCR accuracy, further to integrate these solutions with RPA system.

Education

PH.D | COMPUTER VISION | DEC 2016| IIIT, ALLAHABAD, INDIA

M.TECH | INFORMATION SECURITY | JUNE 2011 | KIIT UNIVERSITY, BHUBANESHWAR, INDIA

M.SC | INFORMATION TECHNOLOGY | JUNE 2009 | KUMAUN UNIVERSITY, ALMORA, INDIA

B.SC | MATHEMATICS | JUNE 2007 | KUMAUN UNIVERSITY, ALMORA, INDIA

Skills & Abilities

- Machine Learning
- Computer vision
- Natural language processing
- Generative AI, LLM, Vision Language Model
- Machine learning operations (MLOps)
- Certified scrum master
- AWS, GCP, E2E cloud platforms
- Python, TensorFlow, Pytorch, MySQL

- Genetic Algorithms, Fuzzy Systems
- Human-Robot Interactions (Nao, Pepper)
- Git, CI/CD, Dockers, Micro Services (REST), JIRA
- IOT – ESP32, MQTT, Jetson Nano, Raspberry, NXP

Certifications

- Machine Learning Operations (MLOps)
  Coursera – Deeplearnining.ai
  Focus Areas – ML Workflows, TFX, Model Deployments and Tracking for Production
  May 3, 2022
- TensorFlow 2 for Deep Learning
  Coursera – Imperial College London
  Focus Areas – TensorFlow 2 APIs, Customized Model Training, Probabilistic Deep Learning
  Jul 9, 2023
- Machine Learning and Soft Computing
  Indian Statistical Institute, Kolkata, India
  Focus Areas – Artificial Neural Network, Computer Vision and Genetic Algorithms
  Dec 20, 2012
- Building Cloud Computing Solutions at Scale
  Coursera – Duke University
  Focus Areas – Containers, Cloud APIs, DevOps, AWS an Azure
  Nov 6, 2023
- Scrum Master Certification Specialization
  Coursera – Learn Quest
  Focus Areas – Agile Methodology and Product Management
  Feb 17, 2023
- Prompt Engineering for ChatGPT
  Vanderbilt University
  Focus Area – Few Shot Examples, Meta Language Creation Pattern, Different prompts patterns
  July 9, 2023

Achievements

- SOPHIA – A H2020 EU PROJECT
  Secured three years post-doctoral position on deep learning perception for human-robot collaboration at LIRMM Montpellier.
- POST-DOCTORAL FUNDING FROM KEMPHE FOUNDATIONS, SWEDEN
  Received 2 years of research funding to pursue my postdoctoral work in Human-Robot Interaction at Umeå University, Sweden.
- 1ST POSITION IN M.SC
  Secured first position in M.Sc for consecutive 2 years at university level.
- MARIE SKŁODOWSKA CURIE ACTIONS (MSCA) 2020
  Applied for the MSCA post-doctoral funding in association with Universitat Rovira i Virgili, Spain and received 85% marks.
- MHRD, INDIA PH.D FELLOWSHIP
  Received 2 years of junior research and 2 years of senior research funding from MHRD for the Ph.D program under robotics and AI lab, IIIT Allahabad.

Projects

FINETUNENING OF LLM FOR Q&A ON WEBSITE DATA

In today's world, simply having an FAQ section or static information on a website isn't enough. We can harness this data to build a conversational AI system that provides more effective and intuitive search and information retrieval. To tackle this, we fine-tuned the LLaMA model using data scraped from the website. Initially, we generated question-answer pairs using the website content and ChatGPT. Once we had sufficient data, we used it to fine-tune the model. The model was then deployed as a conversational AI system using OpenWebUI, enabling users to engage with the content in a more interactive manner.

Techniques Used: Llama, OpenWebUI, LoRA, Data Preparation, Tokenization and Encoding.

FINANCE LARGE LANGUAGE MODEL

We fine-tuned BERT, an LLM model, to handle complex reasoning tasks found in financial reports. We utilized the FinQA dataset, which consists of earnings reports from S&P 500 companies spanning from 1999 to 2019. These reports, typically in PDF format, include multiple pages of financial data, often presented in tables and text. The model generates the mathematical equation necessary to conduct the calculation and further execute that equation on table selected values to get the output value (execution value). The model achieved an execution accuracy of 65.05%.
Techniques Used: BERT, Language Embedding, Program Accuracy, Execution Accuracy

SMART VISION – BRANE ENTERPRISES

We created a smart eyewear prototype equipped with an integrated camera to aid visually impaired individuals. Our in-house fabrication of the PCB board enables essential functionalities like Bluetooth connectivity, USB charging, and live-streaming to an Android device. Additionally, we developed a companion app providing features such as object detection, scene analysis, navigation assistance, and person counting. My role involved collaborating with the team to develop and integrate deep learning solutions into the app. Together, we successfully implemented multiple models for object detection, image/dense captioning, currency recognition, OCR, and more.
Techniques Used: Faster-RCNN, MobileNetV2, CNN+LSTM, YOLO

INCREMENTAL FACE LEARNING – BRANE ENTERPRISES

We developed an Android application for conducting face recognition using transfer learning techniques. We utilized the FaceNet model as the backbone network and added a classification layer for face classification. One challenge with neural networks is catastrophic forgetting, where training the entire network (excluding the backbone) on N+1 classes can lead to performance issues. To address this, we adopted an incremental learning approach that helps to train and deploy the model in 30 seconds reducing the onboarding time.
Techniques Used: Transfer Learning, FaceNet, Incremental Learning

ACTION RECOGNITION – MONTPELLIER UNIVERSITY

Given one second long measure of the human's motion, the system can determine human action. The originality lies in the use of joint angles, instead of cartesian coordinates. This design choice makes the framework sensor agnostic and invariant to affine transformations and to anthropometric differences. On AnDy dataset, we outperform the state of the art classifier. Furthermore, we show that our system is effective with limited training data, that it is subject independent, and that it is compatible with robotic real time constraints. In terms of methodology, the system is an original synergy of two antithetical schools of thought: model based and data-based algorithms. Indeed, it is the cascade of an inverse kinematics estimator compliant with the International Society of Biomechanics recommendations, followed by a deep learning architecture based on Bidirectional Long Short Term Memory.
Techniques Used: Bi-LSTM, Mocap, OpenPose, CNN

VISUAL GROUNDING – UMEA UNIVERSITY

For robots to engage with humans in real-world situations or with objects, they must develop a mental representation ("state of mind") that a) reflects the robots' perception and b) ideally aligns with human comprehension and ideas. Using table-top scenarios as an example, we propose a framework for generating a robot's "state of mind" by identifying the objects on the table along with their characteristics (color, shape, texture) and spatial relationships to one another. The robot's view of the scene is depicted in a dynamic graph where object attributes are translated into fuzzy linguistic variables that correspond to human spatial concepts. This endeavor involves creating these graph representations through a combination of low-level neural network-based feature recognition and a high-level fuzzy inference system.
Techniques Used: Fuzzy Inference System, Mask-RCNN, CNN, Local Binary pattern (LBP), Multi-Layer Perceptron

Talks and Presentations

- Title: An empirical review of calibration techniques for the Pepper humanoid robot's RGB and depth camera.
  Venue: Intelligent systems and application, 5th Sep 2019, London, England.
  Occasion: Presented conference paper in IntelliSys 2019.
- Fusion of gesture and speech for increased accuracy in human robot interaction.
  Venue: 25th International conference on methods and models in automation and robotics, 24th Aug 2019, Międzyzdroje, Poland.
  Occasion: Presented conference paper in MMAR 2019.
- Conflict Detection and Resolution in Table Top Scenarios for Human-Robot Interaction.
  Venue: Computing science department, Umea University, 18th Jun, 2019, Umea, Sweden.
  Occasion: Poster presentation in 31st Swedish AI Society Workshop.
- Deep learning and its applications.
  Venue: UFBI department, Umea University, 15th Jun 2018, Umea, Sweden.
  Occasion: Invited as a speaker at Umea center for Functional Brain Imaging (UFBI) day.
- Sketch drawing by NAO humanoid robot.
  Venue: TENCON a premier international technical conference of IEEE Region 10, 1st Nov, 2015, Macau, China.
  Occasion: Presented conference paper in TENCON 2015.

Selected Publications [Journals and Conferences]

JOURNAL PUBLICATIONS

- Singh, A. K., Baranwal, N., Richter, K. F., Hellström, T., & Bensch, S. (2020). Verbal explanations by collaborating robot teams. Paladyn, Journal of Behavioral Robotics, 12(1), 47-57.
- Singh, A. K., Baranwal, N., & Nandi, G. C. (2019). A rough set based reasoning approach for criminal identification. International Journal of Machine Learning and Cybernetics, 10, 413-431.
- Baranwal, N., Nandi, G. C., & Singh, A. K. (2017). Real-Time Gesture–Based Communication Using Possibility Theory–Based Hidden Markov Model. Computational Intelligence, 33(4), 843-862.
- Baranwal, N., Singh, A. K., & Nandi, G. C. (2017). Development of a framework for human–robot interactions with Indian sign language using possibility theory. International Journal of Social Robotics, 9, 563-574.
- Singh, A. K., Baranwal, N., & Nandi, G. C. (2017). Development of a self reliant humanoid robot for sketch drawing. Multimedia Tools and Applications, 76, 18847-18870.

CONFERENCE PUBLICATIONS

- Singh, A. K., Adjel, M., Bonnet, V., Passama, R., & Cherubini, A. (2022, November). A framework for recognizing industrial actions via joint angles. In 2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids) (pp. 210-216). IEEE.
- Kumar Singh, A., Baranwal, N., & Richter, K. F. (2020). A fuzzy inference system for a visually grounded robot state of mind. In ECAI 2020 (pp. 2402-2409). IOS Press.
- Singh, A. K., Baranwal, N., & Richter, K. F. (2020). An empirical review of calibration techniques for the pepper humanoid robot's RGB and depth camera. In Intelligent Systems and Applications: Proceedings of the 2019 Intelligent Systems Conference (IntelliSys) Volume 2 (pp. 1026-1038). Springer International Publishing.
- Singh, A. K., Chakraborty, P., & Nandi, G. C. (2015, November). Sketch drawing by nao humanoid robot. In TENCON 2015-2015 IEEE Region 10 Conference (pp. 1-6). IEEE.

Please find the complete list of publication at my Google Scholar Profile.

Online Presence

Github: https://github.com/robaita
LinkedIn: https://fr.linkedin.com/in/dr-avinash-kumar-singh-2a570a31