Ryan Salah

(530) 803-8154 | ryansalah53@gmail.com | Towson, MD 21286

Experience

US AI May 2023 – Present

AI Team Lead: NLP — Computer Vision — Data Science *Lanham, MD*

- Led a team of machine learning professionals, providing strategic guidance and technical mentorship to drive innovative solutions and collaborative efforts across multiple AI domains.
- Designed and developed advanced machine learning models for mission-critical applications, ensuring high performance and reliability with comprehensive testing and validation frameworks.
- Created and optimized end-to-end model pipelines, streamlining data preprocessing and model deployment processes, improving operational efficiency by 40% through automated workflows.
- Implemented custom training frameworks and fine-tuning strategies for transformer models, YOLO architectures, and computer vision models to achieve domain-specific performance improvements.
- Spearheaded fine-tuning of large language models and vision transformers using transfer learning techniques, achieving significant accuracy improvements on specialized datasets and business-specific use cases.
- Automated the annotation process through development of intelligent labeling systems and active learning frameworks, reducing manual annotation effort by 70% while maintaining data quality and consistency.
- Championed the design of robust APIs and microservices architecture, enhancing product accessibility and ensuring seamless integration with external systems and platforms for scalable AI solutions.
- Managed the successful deployment of machine learning solutions in production environments, optimizing real-time processing capabilities and reducing operational costs through efficient resource utilization.

Horizon Technology Jun 2020 – Apr 2023

Sr.AI Engineer *Framingham, MA*

- Served as a Senior AI Engineer, responsible for leading end-to-end machine learning projects while providing strategic guidance and mentorship to a dedicated team of professionals.
- Managed and supervised data creation, preprocessing, model training, and deployment for a variety of applications, ensuring data quality and model effectiveness.
- Oversaw extensive data handling, including data collection, cleaning, and annotation, to support the development of high-performing machine learning models.
- Designed and implemented robust data preprocessing pipelines, optimizing feature extraction and data augmentation techniques for enhanced model performance.
- Led the fine-tuning and optimization of models, with a particular focus on object detection, resulting in improved object recognition and tracking.
- Managed the deployment of machine learning solutions in production environments, collaborating cross-functionally to guarantee seamless integration and real-time processing.

Aimfox IT Solutions Jul 2019 – Mar 2020

Mid-level ML Developer *Pompano Beach, Fl*

- Responsible for developing and executing test plans, identifying defects, debugging issues, developing automated test scripts, and documenting the testing process to ensure that hardware or software products meet specified requirements and are free from defects.
- Focused on strong analytical skills, attention to detail, and the ability to collaborate effectively with a team.

CapitalOne Aug 2017 – May 2019

Python Developer *San Francisco, CA*

- Worked as a Python Developer, contributing to the development of software components with a focus on verification and high-level programming design.
- Collaborated with experienced team members to design and implement software modules, ensuring they meet quality and reliability standards.
- Assisted in the creation of automated verification scripts and tools to streamline the testing and validation processes.
- Participated in code reviews to learn and improve software quality and coding practices.
- Contributed to the design and architecture of software systems, applying theoretical knowledge gained during education. Worked on algorithm implementation and data processing, applying problem-solving skills to real-world challenges.

PROJECTS

Enterprise RAG Intelligence Platform | *Python, LangChain, ChromaDB, FastAPI, Transformers*

- Developed comprehensive RAG (Retrieval-Augmented Generation) system capable of processing both structured and unstructured enterprise data including PDFs, Word documents, spreadsheets, databases, and web content for intelligent query responses.
- Implemented multi-modal data ingestion pipeline with support for text extraction from various document formats, table parsing, image OCR, and real-time data streaming from enterprise databases and APIs.
- Built advanced vector embedding system using sentence transformers and custom fine-tuned models, creating high-dimensional semantic representations with contextual understanding for accurate document retrieval and similarity matching.
- Designed sophisticated retrieval mechanisms with hybrid search combining dense vector similarity, sparse keyword matching, and semantic re-ranking to ensure optimal context retrieval from large knowledge bases exceeding 10TB of data.
- Created intelligent chunking strategies with overlapping windows and semantic boundary detection, preserving context integrity while optimizing chunk size for transformer model limitations and retrieval accuracy.
- Integrated multiple LLM backends including GPT-4, Claude, and locally deployed Llama models with dynamic model selection based on query complexity, cost optimization, and response time requirements for enterprise scalability.
- Implemented advanced prompt engineering with context compression, query expansion, and multi-shot learning to improve answer quality and reduce hallucination while maintaining factual accuracy from source documents.
- Built enterprise-grade API with authentication, rate limiting, audit logging, and real-time analytics dashboard providing insights into query patterns, model performance, and system utilization metrics for operational monitoring.

WellB - AI-Powered Hotel Management System | *RAG, Llama, OpenAI*

- Developed comprehensive AI chatbot "WellB" using RAG pipeline with local deployment of Llama and DeepSeek models, integrated with OpenAI GPT for enhanced conversational capabilities.
- Implemented intelligent hotel booking system with multi-model AI approach, enabling natural language booking management through locally deployed and cloud-based LLMs.
- Built emergency alert system with real-time notifications and intelligent priority classification using hybrid AI model architecture.
- Created AI-driven contact management with automated data extraction and contact addition using local Llama model deployment for data privacy.
- Developed booking edit functionality with conflict detection and automated rescheduling using DeepSeek model for complex reasoning tasks.
- Integrated multiple AI services creating cohesive WellB platform with local and cloud-based model orchestration for optimal performance and cost efficiency.

IntelliReport - AI-Powered Report Generation | *Python, NLP, Transformers, FastAPI*

- Developed comprehensive AI-powered report generation system with automated text correction, topic extraction, and intelligent summarization capabilities using transformer-based models and BERT architectures.
- Implemented advanced topic extraction engine utilizing BERT and custom NLP models to automatically identify and categorize key themes, subjects, and concepts from unstructured text data with hierarchical clustering.
- Built intelligent summary generation system using extractive and abstractive summarization techniques to create concise, coherent summaries while preserving critical information and maintaining contextual coherence.
- Created topic-wise summary functionality that segments documents by identified topics and generates targeted summaries for each category, enabling focused analysis and better content organization with semantic similarity scoring.
- Implemented smart text highlighting system that automatically identifies and emphasizes important sentences, key phrases, and critical data points using attention-based models and importance scoring algorithms with confidence thresholds.
- Deployed multi-format report generation with markdown and HTML rendering for direct frontend display and PDF download functionality for seamless user experience across different platforms and devices.
- Built scalable API architecture supporting batch processing with real-time progress tracking and efficient memory management for large document handling and concurrent user requests with load balancing capabilities.
- Integrated advanced analytics dashboard providing insights into document processing metrics, user engagement patterns, and system performance optimization recommendations for continuous improvement.

Smart Media Monitoring Application | *Python, PyTorch*

- Led a comprehensive media monitoring project, involving the development and deployment of multiple AI models, including Speech to Text conversion, sentiment analysis, summary generation, topic modeling, facial recognition, and OCR.
- Worked with both Urdu and English datasets to ensure language versatility and provide valuable insights for the application's users.
- Developed a state-of-the-art Speech to Text model to transcribe spoken content into text, enhancing user accessibility and content indexing.
- Implemented sentiment analysis models to extract and categorize user sentiment from media content, enabling real-time monitoring of public opinion.
- Designed and optimized summary generation algorithms to distill lengthy media articles into concise, informative summaries, improving content consumption.
- Utilized topic modeling techniques to automatically categorize and organize media content, enhancing content discoverability and user experience.
- Pioneered facial recognition solutions to identify and track individuals in media content, providing valuable insights for security and analytics.
- Developed comprehensive text detection and OCR pipeline supporting multi-format document processing including PDFs, images, and tables with multi-language capabilities for Urdu, English, and Arabic text extraction.
- Implemented advanced OCR system with preprocessing techniques for noise reduction, skew correction, and layout analysis to handle complex document structures and varying image qualities.
- Integrated a YouTube Vlogger Scrapper, automating data extraction for content analysis and audience insights.

Smart Surveillance System | *Python, PyTorch and TensorFlow*

- Led the Smart Surveillance project focused on live camera streams and IP cameras, specializing in AI modules for comprehensive security monitoring.
- Live Object Detection: Implemented real-time object detection models for enhanced security and tracking with high accuracy and low latency performance.
- ANPR (Automatic Number Plate Recognition): Developed a highly accurate ANPR system for vehicle monitoring and identification with multi-language support.
- Geofencing: Created geofencing algorithms to define and enforce geographic boundaries within the camera streams for automated area monitoring.
- Video Summarization: Designed video summarization techniques to condense lengthy footage into concise and informative summaries for efficient review.

Interview Process Automation | *Python, FastAPI, PyTorch*

- Led the Interview Process Automation project aimed at automating candidate selection by implementing various AI modules for comprehensive candidate assessment.
- Face Match: Developed a face recognition system to verify candidate identities with high accuracy and anti-spoofing capabilities.
- Face Consistency: Ensured facial consistency during different interview stages to prevent impersonation and maintain interview integrity.
- Voice Consistency: Implemented voice analysis to verify vocal consistency and prevent voice fraud using advanced audio processing techniques.
- Communication Skill Rating: Utilized AI to assess and rate candidate communication skills during interviews with natural language processing.
- Emotion Recognition: Implemented emotion recognition models to gauge candidate emotional responses during interviews for comprehensive behavioral analysis.

AAQAB Project | *Python, Yolo, PyTorch*

- Led the AAQAB Project, utilizing YOLO (You Only Look Once) and ResNet for object detection in large images with focus on satellite imagery analysis and geospatial intelligence applications.
- Large Image Processing: Worked on processing large images and splitting them based on latitude and longitude coordinates for efficient computational handling and distributed processing across multiple GPU clusters.
- Object Detection with SAHI: Implemented object detection algorithms with a focus on accurate object counting in high-resolution satellite imagery using sliding window approach for enhanced detection accuracy.
- PRSS Image Pipeline: Developed a comprehensive pipeline for PRSS (Pakistan Remote Sensing Satellite) image analysis with automated processing workflows and quality assurance mechanisms for consistent results.

- XML to Text Conversion: Automated the process of converting XML annotations to text format, facilitating data preprocessing for object detection models with validation and error handling capabilities.
- YAML Generation for YOLO: Automated the creation of YAML configuration files for YOLO model training, streamlining the training process with parameter optimization and hyperparameter tuning support.
- Daily Report Generation: Automated daily report generation to provide insights based on detected objects with statistical analysis, visualization dashboards, and trend monitoring for operational decision-making.
- Performance Optimization: Implemented GPU acceleration and memory optimization techniques to handle large-scale satellite imagery processing with reduced processing time and improved throughput efficiency.

Crowd Counting Automation | *Python, PyTorch*

- Contributed to a project focused on crowd counting using computer vision techniques to address the challenge of accurately counting people in densely populated areas such as sports stadiums and large-scale festivals.
- Crowd Counting Technique: Collaborated in the development and implementation of crowd counting techniques to accurately determine the number of people in images using density estimation methods.
- Dense Crowd Challenges: Assisted in addressing the unique challenges of counting people in densely crowded scenarios, working towards innovative solutions for occlusion handling.
- Efficiency and Speed: Contributed to designing and optimizing the project for efficient crowd counting in large images with a focus on real-time processing capabilities.
- Practical Implementation: Played a role in translating research findings into practical solutions for automating crowd counting tasks, providing valuable insights in crowded environments.

Smart Meter Reader | *Python, FastAPI, PyTorch, Android*

- Designed and developed the ”Smart Meter Reader” as the final year project in computer engineering to automate the reading of utility meters using computer vision and deep learning techniques.
- ROI Detection: Implemented deep learning models for detecting Regions of Interest (ROI) within meter images, identifying two key areas - the meter digit area and the reference number.
- Image Processing Pipeline: Developed an image processing pipeline that cropped and isolated the selected ROIs for further analysis with noise reduction and enhancement techniques.
- Reading Extraction: Utilized deep learning models to extract readings from the selected ROI areas with high accuracy and error correction mechanisms.
- Android Application Integration: Integrated the project with an Android application that allowed users to upload meter images and receive the automated readings.
- Cross-Network Access: Ensured that the machine-learning model could be accessed from different networks, enabling remote usage and scalability.

Education

University of Rochester Aug. 2013 – May. 2017
Bachelor of Science in Computer Science *Rochester, NY*

Achievements & Awards

Problem Solver Award
Recognized for innovative solutions to complex AI challenges and technical problems *2024*

Publications & Research

Meter Digit Recognition Via Faster R-CNN IEEE
Lead author - Research paper on automated meter reading using deep learning (IEEE Document ID: 8967357) *2019*

Technical Skills

Programming Languages: Python, Flask, FastAPI, JavaScript, TypeScript
IDEs/Development Environments: VS Code, Jupyter Notebook, Google Colab, Anaconda, PyCharm
Libraries/Frameworks: TensorFlow, PyTorch, scikit-learn, Keras, OpenCV, Pandas, NumPy, FastAPI, Transformers, LangChain, LangGraph, Ollama, Hugging Face
Databases: PostgreSQL, MySQL, ChromaDB, Pinecone, Weaviate, FAISS, Qdrant
Development Tools: Jira, GitHub, Git, Docker