(650) 283-3714
manish.sapkota@gmail.com
LinkedIn
Google Scholar

Manish Sapkota

HIGHLIGHTS

- Skilled professional able to develop projects from R&D to full production level with 8+ years of experience in the field of machine learning, computer vision, image processing and natural language processing and 11+ years of experience in the field of software application development.
- Extensive experience in developing and deploying deep learning-based methods in production to solve problems in computer vision (e.g., detection and tracking) and natural language processing, search, and recommendation in scale. Experience with large language models and knowledge distillation.
- Experience in ML backend for training and evaluation and optimizing and scaling these processes. Most of the modeling experience in ***pytorch*** and ***tensorflow*** and deployment experience in ***AWS cloud infrastructure and tools***.
- Not hesitant to do dirty work and able to handle and develop full data pipeline tools to cleanup data, gather gold standards and generate statistics.
- Good in collaboration and teamwork and able to lead the team to solve complex problems.

SKILLS

Scientific & Production: python, c++
Cloud: AWS ecosystem
Machine Learning: pytorch and tensorflow

EMPLOYMENT

Software Engineer, ML **Cresta AI, CA** **Jul 2021 – Nov 2022**

- Envisioned, prototyped, developed, and deployed knowledge base document ingestion, management and semantic search and recommendation system within a short time span of 3 months. Search and recommendation are implemented to understand the conversation context. System is currently part of the overall product suite that serves contact centers and deployed to all the Cresta's enterprise customers.
- Improved retrieval and ranking model to increase relevance of recommended knowledge articles by more than 20% on F1@3 score. Analyzed overall perceived latency and backend latency and reduced it to less than 150 ms to improve the search and recommendation experience.
- Regularly completed opportunity analysis and improvement of the machine learning models for the agent facing conversation context aware features.
- Extensive backend work to enable training to deployment of language models in scale.

Software Engineer, ML Platform **Uber ATG (Acquired by Aurora in 2021), CA** **Feb 2020 – Jul 2021**

- Researched, prototyped, and developed tools to scale deep learning-based methods offline testing to solve computer vision perception and prediction problems and improve the safety of the self-driving cars.
- Improved and integrated overall continuous training, evaluation and deployment pipeline into the end-to-end framework to enable traceability and versioning. Work is used in a broader framework of landing checklists used by model owners for model deployment. Collaborated and helped customers onboard their models to the ML pipeline.
- Collaborated with customers and explored, documented, compared, and recommended different viable ML solutions. Decision is broadly dependent on all the stakeholder’s agreement, cost, speed of development, infrastructure need and availability of support.

Machine Learning Engineer **Foresight AI, CA** **Nov 2018 – Dec 2019**

- Developed deep learning-based methods for 2D object detection and semantic segmentation in videos. Improved detection and instance segmentation performance by more than 40%.
- Implemented most of the machine learning backend in tensorflow and pytorch to train the deep models. Bench-marked a few of the SOTA methods and made changes to the architecture and loss functions to adapt to the Foresight AI data to improve performance and latency.
- Profiled different layers of the pytorch based deep models to understand the bottle necks and implemented changes to improve the overall throughput.
- Implemented scalable framework inference, using multi-process and cloud infrastructures (AWS and Google Cloud). Reduced inference time of 5 minutes 4K video to 30 mins (i.e. 600% improvements).
- Implemented closed loop ecosystem with active learning for data selection, model training, and evaluation.
- Developed software libraries to create generic and clean data consumption pipeline.

EDUCATION

PhD, Computer Engineering **University of Florida** **Aug 2014 – Aug 2018**

- **Thesis:** High Throughput Biomedical Image Analysis and Imaging Informatics for Computer Aided Diagnosis
- Continuation of PhD started at University of Kentucky in 2011.
- Researched, implemented, and evaluated machine learning algorithms, both traditional and state-of-the-arts deep learning methods, to solve the problem of:
    - Object detection and segmentation in images.
    - Feature representation and classification of images and sentences.oBinary representation for fast images/sentences search and retrieval.
    - Computer-Aided Diagnosis.
- Created, implemented and maintained wrapper class to abstract most of the pytorch boiler-plate code and enable faster training and iteration of deep learning based methods for images and natural language understanding.
- Actively implemented new and emerging SOTA deep learning based methods from papers ideation to the working code for comparative studies.
- **Selected Graduate Coursework:** Advanced Machine Learning, Machine Learning, Pattern Recognition, BigDataEcosystems, Cloud Computing, Algorithm Design.