Prasanth Yadla

528 Pontius Ave N, #108, Seattle, WA, 98109
(949) 294-2139 pyadla2@alumni.ncsu.edu

Education

North Carolina State University *Raleigh, NC, USA*
Master of Science in Computer Science, GPA 3.74/4.0 *08/14/2019 - 12/4/2020*
Relevant Coursework - Natural Language Processing, Neural Networks & Deep Learning (Computer Vision), Software Engineering, Automated Learning & Data Analysis, Algorithms for data guided business intelligence, Quantum Computing

Birla Institute of Technology and Science, Pilani *Pilani, India*
Master of Science in Physics, GPA 8.67/10.0 (Dual Degree) *08/05/2013 - 07/09/2018*
Bachelor of Engineering in Computer Science, GPA 8.67/10.0 (Dual Degree)
Relevant Coursework - Machine Learning, Information Retrieval, Data Structures and Algorithms, Design and Analysis of Algorithms, Object Oriented Programming, Operating Systems, Computer Networks, Statistical Mechanics, Quantum Physics, Theory of Relativity

Research Interests

Natural Language Processing, Multimodal Deep Learning, Speech Processing

Work Experience

Apple Inc. **Seattle, WA, USA**
Senior Machine Learning Engineer *07/24/2024 - Present*

- Develop deep learning layers and models which are compatible with core-ml-tools and Apple Neural Engine (ANE). Research and design audio encoder and decoder layers for Automatic Speech recognition on-device in Siri.
- Research, design and develop large-language models and vision models for **Apple Intelligence** and **Siri Automatic Speech Recognition**. Palletize and quantize layers effectively to reduce memory footprint. Various specific use-cases for Apple Intelligence include email Summarization, urgency classification, etc.
- Co-develop and maintain the library for distributed training and evaluation. Enhance the performance of training algorithms like Fully-Sharded Data Parallel & Distributed Data Parallel.
- Develop and maintain serving infrastructure for large-language models, deploying models more than 65 billion parameters in size.

Apple Inc. Seattle, WA, USA
Machine Learning Engineer *04/24/2023 - 07/23/2024*

- Built a library containing state-of-the-art speech and audio layers and models for a variety of use-cases like automatic speech recognition, text-to-speech,etc which are fully-optimized to run on On-device neural engine chips. Built a novel speech recognition conformer encoder with innovative positional encoding techniques achieving 1.5x speedup in inference latency and 1.2x speedup in training with 15% improvement in memory footprint.
- Designed and implemented various model checkpoint conversion algorithms for exporting state-of-the-art foundation models to compressed and optimized formats to deploy On-Device (Apple Neural Engine) powering Apple devices and Server-side inference powering Apple subscription products including Apple Music, Podcasts, AppleTV, etc.
- Co-Develop and maintain a performant platform for large-scale distributed training capable for pre-training and evaluating multimodal foundation models greater than 65 billion parameters in size.
- Help product teams fine-tune foundation models for downstream use-cases like Siri Summarization, Siri Automatic Speech Recognition and Machine Translation
- Co-develop and maintain a machine learning fine-tuning library being used for various use-cases like Apple Pay logo detection, Street sign detection in Apple Maps, Accessibility in iOS, Scene classification in Apple camera,etc.

Amazon.com Services LLC **Irvine, CA, USA**
Software Development Engineer *02/01/2021 - 04/21/2023*

- As part of the personalization team in Amazon Music, responsible for the Machine Learning platform providing real-time personalized ranked list of live radio shows for 1M+ listeners of the Amp App. Technologies used - Redis, AWS SageMaker, PySpark, ECS Fargate, Kotlin
- As part of the personalization team in Amazon Music, Developed & Deployed an API for show and creator recommendations interfacing with SageMaker endpoint for inference. Built a real-time and batch pipeline for computing point-in-time features and affinities between listeners and creators.
- As part of the Alexa AI Web Info team, responsible for the big data platform to crawl, store & process petabyte-scale web content data powering search queries of various domains in Alexa via question answering model. Technologies Used - Scala, Spark, Airflow, Delta Lake, AWS Lambda, S3
- Built a lexical and neural index for retrieving top-K documents given a query which is fed to the Reranker module and ultimately passed through a RoBERTa variant language model for inference of candidate sentences.
- Designed and Implemented cost-effective Backup Strategy for large-scale Delta Lake, taking the latest snapshot of data files and transaction log, producing 60% cost savings.
- Designed & implemented an automated Vacuuming strategy to maintain rolling window of data files, deleting older files greater than retention threshold of 21 days, saving 70% of rolling-window storage cost.
- Designed and developed an asynchronous message processing system using AWS SNS, AWS SQS and AWS Lambda for Alexa Alarm Tones release in the US. Currently processing requests at 100 Transactions per day.
- Designed and developed a real-time ETL data ingestion pipeline using AWS Kinesis, Firehose, S3, Glue & Redshift for storing events related to digital product purchase metrics within skill sessions for BI purposes. Currently ingesting 1k events/sec.

Amazon.com Services LLC Irvine, CA, USA
SDE Intern *06/08/2020 - 08/14/2020*

- Developed a new feature allowing customers to discover more In-Skill Products in Headed Alexa-enabled devices like Echo Show 5 in the US marketplace, improving digital product discovery & purchasability by 10%. Gained exposure to Alexa Orchestration workflow and multi-modal Alexa Skills development along with VUI design

Oracle India Pvt. Ltd Bangalore, India
Software Engineer *07/26/2018 - 08/02/2019*

- Designed, developed and maintained high performance multithreaded APIs consumed by Oracle Sales Cloud Application UI, currently being used by approximately. 20,000 users. at Oracle
- Adapted Guava Caching Mechanism into DB read Endpoints, resulting in 100x latency improvement, thereby resolving High Severity Production Tickets from Business
- Devised Real-time usage reporting and statistics UI leveraging ElasticSearch, Logstash and Kibana (ELK) Stack

Cloudera Inc. Bangalore, India
Intern *01/08/2018 - 06/18/2018*

- Designed & Developed a data generating library, generating large-scale randomized data in a distributed fashion using Hortonworks Data Platform from a provided config. This data is used for testing purposes for HDP clusters.
- Technologies Used - PySpark, HDP

Amazon India Pvt. Ltd. Bangalore, India
Intern *07/10/2017 - 12/15/2017*

- Re-architectured of inventory service using Hibernate and spring which included data migration from existing amazon dynamoDB to Amazon aurora, reducing latency issues encountered previously
- Developed a new automated workflow execution system for enabling order cancellation in Amazon Seller Flex. Processing 200+ order cancellation requests per second.
- Technologies used - Java, Spring, AWS

IBM India Pvt. Ltd Bangalore, India

Intern *05/26/2015 - 07/25/2015*

- Developed Unit test case suit for MaaS On premises Configuration User Interface Web Application which included Models, Views and Configurations
- Technologies Used - Django, Python

Academic Projects and Research Experience

Data Assistant**,** Industrial and Systems Engineering, NCSU Dec 2019 - April 2020

Under Prof. Julie Swann

- Performed large-scale data transformation of healthcare EHR files containing 15 Million records which is to be processed, aggregated and filtered in a matter of minutes using a 80 core linux machine
- Data from the pipeline built was used to model urinary tract infection hospitalizations using hierarchical clustering & LASSO-logistic regression and was published as a conference paper

Improving Question Answering System Performance using Language Heuristics & Knowledge Distillation

Under Prof. Muninder P Singh, NCSU Aug 2020 - Nov 2020

- Performed data augmentation and linguistic post processing techniques, along with knowledge distillation to improve F1 and EM score, along with latency & performance of standard question-answering tasks with SQUAD 2.0 dataset using baseline BERT model.
- Technologies Used - Tensorflow, PyTorch, Pandas, Numpy, Google Collab on GCP

Plant Phenotyping using Object Detection

Under Prof. Edgar Lobaton, NCSU Feb 2020 - April 2020

- Objective is to apply image processing & CNN to given images to detect leaves and collars for identifying the phenotype features of a plant in automated fashion.
- Trained various models like SSD MobileNet V2, Inception over different hyperparameters & compared inference results with Baseline (MobileNet V1)

Text Based Information Retrieval System Sept 2016 - Nov 2016

Under Prof. Aruna Malapati, BITS Pilani, India

- Developed a basic text-based ranking search engine based on vector space model using TF-IDF weights. Index construction was built using dictionary data structure.
- Technologies Used - Python, NLTK

Research Assistant, Department of CDS, Indian Institute of Science Jan 2018 - July 2018

Under Prof. Yogesh Simmhan

- Wrote an algorithm to schedule a transactional DAG of tasks across a fabric of Cloud, Fog and Edge, that are triggered based on temporal, spatial and domain features of generated inputs, to execute within a deadline.
- Work published as a poster in Robert Bosch Center for Cyber Physical Systems at IISc.

Research Assistant, Intel IoT Lab, BITS Pilani, India

Under Prof. Chittaranjan Hota Jan - May 2017

- Designed an experimental IoT testbed interspersed with sensors streaming raw data over the network.
- Developed an efficient dimensionality reduction algorithm for raw data using the SAX (Symbolic Aggregate Approximation) algorithm and applied K-means and Markov-Based rule mining algorithms to process the data.The system developed is being used for data analysis purposes in the IoT lab.
- Languages and Libraries - Python, R, NumPy, SciPy and Matplotlib

Honors and Awards

INSPIRE Scholarship(SHE) Aug 2013 - July 2018

Govt. of India

This scheme offers Rs.80,000/- every year for students who happen to be in the top 1% in their respective Board Examinations and are pursuing courses in Natural and Basic sciences at the B.Sc. or Integrated M.Sc. levels.

Languages and Frameworks

PyTorch, JAX, Tensorflow, C++, Python, Java, Python, Scala, Git, AWS, Spark, Kubernetes, Docker

Certifications

- AWS Certified Solutions Architect-Associate (Issued Sep 2021)
- Neural Networks and Deep Learning Coursera (Issued Sep 2018)