Aran Khanna

San Francisco, CA http://arankhanna.com

Experience

Amazon Web Services (AWS) Palo Alto, CA
AI Engineer, AWS Deep Learning October 2016-February 2018

- Tech lead for the AWS DeepLens product. Pitched the initial idea to management, drove the hardware partner selection of Intel, defined the architecture and launch content, helped hire the development team of 8 people, coordinated development efforts between Intel, our ODM and Amazon, and led a launch of the product at AWS re:Invent in 6 months. With thousands of units shipped and the product ranked in the top 10 of first party devices on Amazon.com as of Q1 2018.
- Tech lead for deep learning on edge (IoT/mobile) devices as the eight member of the AWS Deep Learning team. Worked with ARM, Intel, Qualcomm, NVIDIA and Apple to coordinate deep learning workload support for the Apache MXNet framework on their chipsets, leading to order of magnitude performance gains. Contributed code to integrate MXNet and AWS IoT services with Apple CoreML, Raspberry Pi and NVIDIA Jetson leading to adoption of these platforms by AWS customers for edge ML workloads.
- Mentored PhD-candidate interns conducting novel research in deep network model compression and efficiency, working closely with professors Zachary Lipton, Anima Anandkumar and Alex Smola. Produced five conference papers and contributed to the development of novel, GPU-backed modeling tools to Apache MXNet such as weight pruning, sum product networks, tensor contraction and tensor regression layers. Presented work at ICLR 2018 (poster), CVPR 2017 (workshop), ICML 2017 (tutorial) and NIPS 2017(workshop).
- Filed eight patents around secure and dynamic deep learning systems for IoT deployments.
- Evangelized MXNet and corresponding AWS AI services within Amazon (Lab126 (Alexa), Amazon retail and AmazonGo) and to global AWS customers (in Europe, Japan, India, U.S.). Wrote AWS AI Blog posts, delivered a numerous talks, including keynotes and sessions at major AI industry and academic conferences.

Harvard Institute For Quantitative Social Science (IQSS) Cambridge, MA
Research Fellow May 2015-August 2016

- Researched and identified major privacy and security vulnerabilities in products used by millions of people worldwide, leveraging technologies such as distributed bot-nets, penetration testing tools and data visualization tools, working closely with Professor Latanya Sweeney to produce two academic papers published in the Journal of Technology Science.
- Spoke and wrote extensively about digital privacy including a talk at TEDxBerkeley, an Op-Ed in TIME magazine, interviews with several global publications including the BBC and Forbes, and regular contributions to the Huffington Post and Business Insider including a popular blog post on Facebook Messenger's location privacy, which was shared on over 200 global news publications with the corresponding Chrome extension being downloaded over 85,000 times.

Marianas Labs (Machine Learning Startup, Acquired 2016) Mountain View, CA
Software Development Intern June 2015-August 2015

- Wrote a large-scale distributed data crawling and aggregation framework in Python that was licensed to Marianas Labs customers for an incremental $200k of revenue.
- Interned at an early stage machine learning startup directly under Ashfaq Munshi, former CTO of Yahoo, and Professor Alex Smola of Carnegie Mellon Machine Learning.
- Developed a clustering method to identify new object classes in terabytes of images using a deep-network-based image embedding system, built in Caffe, leading to a significant reduction in data mislabeling.

Novus Partners (Financial Risk Modeling Startup) New York, NY
Software Development Intern June 2014-August 2014

- Interned alongside a team of engineers to add features to the core platform that consumes hedge fund data and provides quantitative analysis (exposures, position sizing etc.)
- Worked in Scala on the data pipeline building out more efficient data consumption processes, and built visualizations in Angular and D3 for the consumed data, leading to faster client onboarding and a better user experience.

Microsoft Azure Redmond, WA
Software Development Intern June 2013-August 2013

- Interned in the Windows Azure Fabric Fundamentals Team. Identified the causes of infrastructure downtime by developing a C# based data pipeline on top of core VM allocation systems.
- Built a distributed unsupervised classification engine, based on hierarchical clustering methods to analyze and bucket the terabytes of operational data coming from the entire network, helping triage hundreds of bugs and increase VM availability.

Academic Publications

M Tschannen, Khanna A, Anandkumar A. StrassenNets: Deep Learning With a Multiplication Budget. *Poster Presentation and Long Talk*. July 11, 2018. ICML 2018, Stockholm, Sweden.
Dhillon G, Azizzadenesheli K, Khanna A, Bernstein J, Kossafi J, Lipton Z, Anandkumar A. Stochastic Activation Pruning For Robust Adversarial Defense. *Poster Presentation. ICLR* 2018. Vancouver, Canada.
Kossafi J, Lipton Z, Khanna A, Furlanello T, Anandkumar A. Tensor Contraction & Regression Networks. *Best Poster Winner. NIPS MLTrain Workshop*. December 9, 2017. Long Beach, California. ArXiv. Print.
Kossafi J, Khanna A, Lipton Z, Furlanello T, Anandkumar A. Tensor Contraction Layers for Parsimonious Deep Nets. *CVPR Tensor Methods in Computer Vision Workshop*. July 26, 2017. Honolulu Hawaii. ArXiv. Print.
Khanna A. Facebook's Privacy Incident Response: a Study of Geolocation Sharing on Facebook Messenger. *Technology Science*. 2015081101. August 11, 2015.
Khanna A. Venmo'ed: Sharing Your Payment Data With the World. *Technology Science*. 2015102901. October 29, 2015.

Awards and Patents

StrassenNets: Deep Learning With a Multiplication Budget ICML 2018 poster presentation and long talk
Stochastic Activation Pruning For Robust Adversarial Defense, ICLR 2018 poster presentation
Aran Khanna. Method for Achieving Consistency in Distributed Edge Models. US patent pending, 2017
Aran Khanna. Auto Normalization Models to Preprocess Sensor Data. US patent pending, 2017
Aran Khanna. Model Deployment Across IoT Networks. US patent pending, 2017
Aran Khanna. Secure Models for IoT Devices. US patent pending, 2017
Aran Khanna. Generating Adaptive Models for IoT Networks. US patent pending, 2017
Aran Khanna; Calvin Kuo; Sunil Mallya. Split Predictions for IoT Devices. US patent pending, 2017
Calvin Kuo; Sunil Mallya; Aran Khanna. Training Models for IoT Devices. US patent pending, 2017
Calvin Kuo; Aran Khanna; Sunil Mallya. Model Tiering for IoT Device Clusters. US patent pending, 2017

Technical Skills

Software: AWS, Azure, Tensorflow, Pytorch, MXNet, DMLC, CoreML, Swift, ObjectiveC, Linux, C++, Java, C#, Scala, Python, Javascript, React, D3, Jenkins, HTML/CSS, SQL, Mongo, Docker
Language: Proficient in Spanish

Education

Harvard University Cambridge, MA
A.B. in Computer Science, Secondary in Mathematical Sciences with Concentration GPA of 3.8 August 2012 –May 2016
Relevant coursework: Machine Learning, Data Structures and Algorithms, Operating Systems, Topology I, Number Theory, Combinatorics, Honors Linear Algebra and Real Analysis, Computational Complexity in Biology, Theory of Computation, Systems Security, Programming Abstraction and Design, Distributed Systems, HBS Economics of Engineering, HBS Innovation and Entrepreneurship, Design of Usable Interactive Systems, Privacy and Technology, Public Speaking

Lakeside High School Seattle, WA
Graduated with a GPA of 3.9 August 2008-June 2012
AP Scholar with Distinction, National Merit Finalist, 2012 Captain of Varsity Rowing Team, 2012 Managing Editor of School Paper