Apurva Patkeshwar

apurva.p@hotmail.com • apurva.dev • 425-375-8654

Experience

Twitter Seattle, WA
Machine Learning Engineer Jul. 2021 – Present

- Working on the next-gen fully-managed ML Serving service in Cortex platform which will enable other teams to serve ML models at high scale in production while ensuring performance, reliability, and ease of use [Scala]

Amazon Seattle, WA
Software Development Engineer Jun. 2018 – Jul. 2021

- As a part of Alexa Speech Recognition (ASR) team, worked on separating audio streaming path from rest of the architecture and cache audio in regions closer to customers. This brought speech processing to edge of the cloud reducing user perceived latency (UPL) by 25% at 99th percentile in emerging markets [Java, Rest API]
- Led the migration from push-to-talk to wake-word models, by turning off end-pointing in cloud run-time. Re-routed 1 million daily requests to ensure a smooth transition reducing operational overload and shortening overall rebuild execution time by 9 hours. This saved model release cost by 15% ($500K/year), model rebuild cost by 25% ($170k/year) and $120k/year in build-time computation cost for benchmarking [Java, XML]

Microsoft Hyderabad, India
Software Engineer Jul. 2014 – Jul. 2016

- Developed a real-time web app to visualize occupancy, noise and temperature data about office spaces, fetched from custom 3D printed sensors. Built functionality to reserve conference rooms with just an RFID badge swipe which led to a 3x increase in monthly usage [C#, jQuery, AJAX, SQL]

Education

Courant Institute of Mathematical Sciences, New York University New York, NY
Master of Science in Computer Science, 3.7/4.0 May 2018

Sardar Patel Institute of Technology, University of Mumbai Mumbai, India
Bachelor of Engineering in Information Technology, 3.6/4.0 Jun. 2014

Projects

Identifying duplicate questions on Quora Sep. – Dec. 2017

- Trained a bidirectional LSTM neural network on an 800k question dataset. Using soft attention for alignment, detected semantically duplicate questions with an accuracy of 85.26%. The questions were pre-trained as 200-dimension vector representations using GloVe word embeddings [NLP, TensorFlow, Keras]

Classifying restaurants using photos on Yelp Sep. – Dec. 2017

- Used transfer learning to extract features from user-uploaded photos by leveraging 4 AlexNet-derived models. Predicted restaurant labels (like good for lunch, outdoor seating, expensive etc.) from these features using a custom deep learning ensemble with an F1 score of 0.825 [Computer Vision, Caffe, Theano, Places205]

Awards

Forbes Under 30 Scholar Oct. 2017

- Selected among 1000 high performing students across US and invited to attend the Forbes Under 30 Summit

Google Code Jam Mar. 2017

- Finished in the top 150 worldwide and was invited to attend the Google I/O 2017 developer conference

Additional Experience

Citi, *ML Engineer Intern* Feb. – May 2018
New York University, *Teaching Assistant* Sep. 2016 – May 2018
Yahoo, *Software Engineer Intern* May – Aug. 2017
Microsoft, *MACH Education Committee Lead* Jan. – Jul. 2016
Microsoft, *Software Engineer Intern* Jun. – Aug. 2013

Skills

Languages: Java, Scala, C#, C++, Python, JavaScript, PHP, MySQL
Platforms: AWS, GCP, Hadoop, PyTorch, TensorFlow, .NET, LaTeX