Raul Puri

34304 Portia Terrace, Fremont, CA 94555 (510)-584-8347 raulpuric@berkeley.edu
https://www.linkedin.com/in/raul-puri-3a0b43a0 https://github.com/raulpuric

Stanford University **January 2019 –**
AI Master's Degree Certificate

Highlighted Coursework

- CS228 – Probabilistic Graphical Models, CS276 – Information Retrieval and Web Search, CS330 - Deep Multi-Task and Meta Learning

University of California, Berkeley **August 2013 – May 2017**
B.S. in Electrical Engineering and Computer Science

Highlighted Coursework

- CS294.129 – Deep Convolutional Neural Network design, CS294.131 Special Topics in Deep Learning, EE127 – Optimization Models and Applications, CS162 – OS and Systems Programming, CS194.15 – Engineering Parallel software, CS189 – Machine Learning, CS188 – Introduction to Artificial Intelligence, CS170 – Efficient Algorithms and Intractable Problems, CS9E – Unix

Experience

Experienced leader, author, and intern mentor in research and development of Artificial Intelligence (AI), Machine Learning (ML), Natural Language Processing (NLP), Robotic Automation, Security, Cloud/High Performance Computing, Systems, and Parallel Algorithms.

NVIDIA – Deep Learning Research Scientist **08/17 –**

- Lead growth of NLP research team as an integral early hire.
- NLP research expertise in Question Answering and Question Generation for Search, Dialogue Modeling and Chatbots, Meta Learning for Language, and Text Classification.
- Systems and HPC research in distributed, data & model parallel, and mixed precision training for State-of-the-Art Neural Language Models (LMs). Using 12 ZettaFLOPs of compute we trained the world's largest LM with 8.3 billion parameters on a top500 supercomputer.
- Production experience training and deploying deep learning APIs for Question Answering and Search, Conversational AI, and classification for public demos and internal use.
- Lead, developed, and released multiple open source projects with over 2000 GitHub stars.

A10 Networks – Senior Machine Learning Engineer **12/16 – 08/17**

- Developed Deep Learning models to detect malicious traffic and model network packet flow of a distributed network. The models developed were able to accurately identify large scale botnet DDOS attacks, the likes of which brought down websites like GitHub and Twitter.

ML@B (ML at Berkeley) – Head of Education, Project Manager **08/16 – 05/17**

- Created curricula for student-taught ML classes on campus, managed a reading group, trained new ML@B members in skills needed to be ML researchers/consultants, edited ML blog posts, wrote and published Deep Learning blog posts with O'Reilly media
- Created new research projects and managed new/existing research and consulting projects

CS189: Intro to Machine Learning – Student Instructor **12/16 – 05/17**

- Taught Machine Learning fundamentals to Berkeley Graduate and Undergraduate students

Robotics Automation Lab – Researcher for Prof. Ken Goldberg **11/14 – 05/17**

- Developed image segmentation system trained on MS COCO/PASCAL VOC2017 for detecting deformable objects in autonomous robotic surgery

UnifyId – Fellow **08/16 – 11/16**

- Researched mitigation of adversarial attacks against DL-based authentication methods

Students Mentored

- Alex Boyd – 2nd year PhD student @ UC Irvine. Conversational Modeling. **06/19 – 01/20**
- Ryan Spring – 6th year PhD student @ Rice University. Question Answering. **05/19 – 09/19**
- Neel Kant – 4th year Undergraduate student @ UC Berkeley. Transfer Learning in NLP. **06/18 – 09/18**

Publications

- **Raul Puri**, Ryan Spring, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro. *Training Question Answering Models From Synthetic Data*. Under review at ICML 2020
- Alex Boyd, **Raul Puri**, Mohammad Shoeybi, Mostofa Patwary, Bryan Catanzaro. *Large Scale Unsupervised Generative Dialog Modeling with Personality Transfer*. Under review at ACL 2020.
- **Raul Puri**, Bryan Catanzaro. *Zero-Shot Text Classification With Generative Language Models*. MetaLearn 2019 @ NeurIPS. http://metalearning.ml/2019/papers/metalearn2019-puri.pdf.
- Mohammad Shoeybi, Mostofa Patwary, **Raul Puri**, Patrick LeGresley, Jared Casper, Bryan Catanzaro. *Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism*. arXiv. 2019. https://arxiv.org/abs/1909.08053
- Neel Kant, **Raul Puri**, Nikolai Yakovenko, Bryan Catanzaro. *Practical Text Classification with Large Pre-trained Language Models*. arXiv. 2018. https://tinyurl.com/practicaltext2018
- **Raul Puri**, Robert Kirby, Nikolai Yakovenko, Bryan Catanzaro. *Large Scale Language Modeling: Converging on 40GB of Text in Four Hours*. HPML: High Performance Machine Learning. 2018. http://arxiv.org/abs/1808.01371 (**Best Paper**)
- Phillip Kuznetsov, Riley Edmunds, Ted Xiao, Humza Iqbal, **Raul Puri**, Noah Golmant, and Shannon Shih. *Adversarial Machine Learning. Artificial Intelligence Safety and Security* (Chapman & Hall/CRC Artificial Intelligence and Robotics Series): Roman V. Yampolskiy, 2018. 235-248. https://drive.google.com/file/d/1OGCI0GGQIADUsYrZPU5BWaE5C4mVkuPl/view?usp=sharing
- Riley F. Edmunds, Noah Golmant, Vinay Ramasesh, Phillip Kuznetsov, Piyush Patil, **Raul Puri**. *Transferability of Adversarial Attacks in Model-Agnostic Meta-Learning*. 2017 Deep Learning and Security Workshop (DLSW) in Singapore. 2017. http://rileyedmunds.com/pdf/dlsw2017.pdf
- **Raul Puri**, Dan Ricciardelli. *Caption This, With Tensorflow*. O'Reilly Media. 2017. https://www.oreilly.com/learning/caption-this-with-tensorflow

Class Projects

CS 330 Deep Multi-Task and Meta Learning
- Applied Model Agnostic Meta Learning to transformer-based extractive QA models to achieve fast adaptation and +3% absolute improvement in SQuAD1.1 score.

CS 294-131 Special Deep Learning NLP Topics Projects
- Designed a variational word embedding algorithm by learning to embed the dictionary.

CS 294-129 Deep Convolutional Neural Net Projects
- Designed a novel video compression algorithm via frame rate upscaling with Variational Auto-Encoders.

CS 194 Parallel Programming Projects
- Implemented scalable GPU parameter sharing support for Tensorflow as detailed by https://papers.nips.cc/paper/4687-large-scale-distributed-deep-networks.pdf
- Implemented Linear Algebra and key data parallel primitives in openMP, pThreads, MPI, and OpenCL, all with CL/SIMD vector support.

CS 189 Machine Learning Projects
- Implemented an MNIST classifier with >99% accuracy. Top 10 in class of 400.

CS 188 AI Projects
- Did multiple projects to beat Pacman >90% of the time including under partial observations, and using raw images fed to a CNN.

CS 170 Algorithms Projects
- Developed an approximation algorithm to solve the NP-Hard Feedback Arc Set for Tournaments problem.

CS 162 Operating System Projects
- Implemented in C: bash shell, HTTP server, Malloc, 2PC/KV store, Unix BSD 4.2 filesystem.
- Combined all of them to make a multithreaded x86 OS.

Experience in Programming and Frameworks/Tools

- Python, Pytorch, TensorFlow, C/C++, Java, OpenCL, pThreads, openMP, MPI, Spark, Most Markup languages (HTML, HTML 5, XML, etc.), CSS, JavaScript

References

- Bryan Catanzaro – VP of Applied Deep Learning Research (bcatanzaro@nvidia.com)
- Rajkumar Jalan – CTO @ A10 Networks (rjalan@a10networks.com)
- Jonathan Shewchuk – Professor & TA employer (jrs@cs.berkeley.edu)