KUSHAL ARORA

Email: kushal18@gmail.com Phone: +352-871-5169
Website: http://kushalarora.github.io
Github: https://github.com/kushalarora

EDUCATION:

Master of Science, Computer Engineering **December 2015**
University of Florida, Gainesville, FL CGPA: 3.74
Courses Taken:
Maths for Intelligent Systems, *Machine Learning*, *Advanced Machine Learning*, *Cloud Computing and Storage*, *Advance Data Structures*, *Analysis of Algorithms*, *Computer Architecture*, *Distributed Operating System*
Master's Thesis:
Compositional Language Modeling (pdf, code)

B. Tech, Electronics and Communication Engineering **May 2010**
Motilal Nehru National Institute of Technology, Allahabad

PUBLICATIONS:

Sachin Grover, Kushal Arora, and Suman K. Mitra. "*Text extraction from document images using edge information.*" *India Conference (INDICON), 2009 Annual IEEE*. IEEE, 2009.

PREPRINTS:

Kushal Arora and Anand Rangarajan. "*A Compositional Approach to Language Modeling.*" arXiv:1604.00100 [cs.CL], 2016. (pdf, code)
Kushal Arora and Anand Rangarajan. "*Contrastive Entropy: A new evaluation metric for unnormlized language models.*" arXiv:1601.00248 [cs.CL], 2016. (pdf, code)

RESEARCH:

Compositional Language Model (pdf, code)
Supervisor: Prof. Anand Rangarajan, University of Florida, Gainesville
Traditional language models treat language as a linear chain on words. In my Master's thesis I challenged this assumption and proposed a model that uses underlying compositional structure for modeling language. This is done by marginalizing the joint probability of sentence and its composition trees. Composition trees were generated using PCFGs and marginalization was carried out using the Inside algorithm. Conditional probability given the structure was modeled using the distributional representation similar to neural network based models. We report more than 100% improvement in Constrastive Entropy over RNNLM on a toy data set.

Contrastive Entropy: A new metric for evaluating unnormalized language models (pdf)
Supervisor: Prof. Anand Rangarajan, University of Florida, Gainesville
Perplexity is an unsuitable metric for sentence level language models due to its word level model assumption and its reliance on exact probabilities. As part of my thesis I also proposed a new discriminative metric to evaluate unnormalized sentence level language models. The intuition here is to capture the model's ability to differentiate between a test sentence and its distorted version. I also hypothesize that this metric will have better correlation with the WER as both metrics are discriminative in nature.

Text Extraction from an Image using Edge Information (pdf)
Supervisor: Prof. Suman K. Mitra, DAIICT, Gandhinagar

In this work we proposed a novel method of marking the text areas in an image. The proposed method was based on collecting the edge information using Sobel operators and then harnessing the property of sharp edges for the text and there by marking the areas as text or non text regions.

PROFESSIONAL EXPERIENCE:

Amazon **September 2015 - present**

***Software Engineering Intern,* Alexa Machine Learning Platform**

Working with Alexa's Machine Learning Platform team on ASR and NLU modeling infrastructure. My current assignment includes designing a pipeline for building supplemental model to support utterances missed by the monolithic static model.

Amazon **May 2014 -August 2014**

***Software Engineering Intern,* Transactional Risk Management Services**

Analyzed counterfeit spike problem for high volume items on third party marketplace. Also designed a generic framework that flags and block sales of dubious products based on a certain criteria.

Chatimity **Sept 2011-June 2013**

Software Engineer

First employee at Chatimity. Worked with two founders to build the complete technology stack from scratch. Helped build a system that handles 30,000 uniques and 11.5 million messages per day (upto June 2013). Designed and launched central product features including topic pages, search, recommendations, notifications and image sharing.

ST-Ericsson **Aug 2010-Sept 2011**

System Software Engineer, Multimedia Audio Team

Developed OpenMaxIL layer components for Audio 3D Mixer and AAC Encoder and implemented features like http streaming, buffering and seek features at framework level.

SELECTED PROJECTS:

Compositional Language Model (code)

Implemented the idea proposed in the paper "*Compositional Approach to Language Modeling*". The code was written in Java. Nd4j was used as main matrix library and UJMP to store Inside-Outside scores due to its support for sparse matrices. For Inside-Outside score calculation the grammar from the Stanford Parser was used. The code was written in a modular way to allow plugging in other grammars without much effort. To my best knowledge this is the first software that does phrasal embedding in a compositional way.

Sentence Level Recurrent Neural Network (code)

Implementation of Sentence Level RNN described in "*Contrastive Entropy: A new evaluation metric for unnormalized language models*". The implementation was done using Theano and Numpy.

Comparative evaluation of Manifold Learning Algorithms (code)

Implemented the state of the art dimensionality reduction algorithms in python using scipy and compared them on four data sets, namely *RaceSpace*, *Digits*, *Faces* and *Swiss Roll*. The project was an individual effort and done as a course project for Advanced Machine Learning class.

Algorithms implemented: *Local Linear Embedding, ISOMap, Laplacian Eigenmaps, Hessian LLE, Local Tanget Space Analysis* and *Stochastic Neighborhood Embedding*

Comparative evaluation of Supervised Learning Algorithms (code)

Built a generic framework to run a list of Supervised Learning Algorithms in Python using Scikit-Learn and Theano. The framework was used to do a comparative study on following data sets: *Wisconsin Breast Cancer*, *Iris*, *Higgs*, *OCR* and *Hand Writing Recognition* across a range of supervised learning algorithms. This project was done in a team of three for Machine Learning class.

Algorithms evaluated were: *Multi Layer Preceptron, Stacked Auto Encoders, Deep Belief Network, Support Vector Machine, Random Forest, Decision Tree* and *AdaBoost Decision Tree*.

Ontology Alignment for Knowledge Bases (code)

Implemented and evaluated PARIS, an ontology alignment technique that uses web text based interlingua for aligning relations and entities. Ontologies for Freebase, NELL and Yago were mapped to each other using label propagation algorithm. This project was done as a independent study in Data Science Lab under Dr Daisy Wang and was a part of larger objective to build a master KB for the lab.

TECHNICAL SKILLS:

Languages: C, C++, Java, Python, Javascript, CSS, MySql, Matlab, Latex, Lyx, Scala

Tools: Git, GDB, MongoDB, Hadoop, Makefiles, Android SDK, Solr, Tornado Web Server, Theano