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 **Aug 2013 - Dec 2015**
University of Florida, Gainesville, FL
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 **July 2006 - 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. (pdf)

PREPRINTS:

Kushal Arora and Anand Rangarajan. "*A Compositional Approach to Language Modeling.*" arXiv:1604.00100 [cs.CL], 2016. (pdf)

Kushal Arora and Anand Rangarajan. "*Contrastive Entropy: A new evaluation metric for unnormalized language models.*" arXiv:1601.00248 [cs.CL], 2016. (pdf)

RESEARCH:

Compositional Language Model (pdf)
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 framework that uses the underlying compositional structure for modeling language. This was done by marginalizing the joint probability of sentence and the composition trees. Composition trees were generated using PCFGs and marginalization was carried out using the Inside algorithm. The conditional probability given the structure was modeled as a distribution over a continuous embedding space.

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 modelling assumptions and its reliance on exact probabilities. As part of my thesis, I proposed a new discriminative metric to evaluate unnormalized language models like sentence-level 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 thereby marking the areas as text or non-text regions.

PROFESSIONAL EXPERIENCE:

Amazon **Aug 2016 - present**

***Software Engineer,* Alexa Machine Learning Algorithms**

Part of the team that develops in-house deep learning library and other ML libraries used at Alexa.

Major contributions:

1. CRF speed up (2.3x improvement with 4 nodes) by implementing MPI-based distributed training.
2. Optimized distributed training for the deep learning library based on [1] for 45% speed up.

I also organize a bi-weekly deep learning reading group for Alexa Machine Learning Platform Group.

Amazon **Sept 2015 - Aug 2016**

***Software Engineer,* Alexa Machine Learning Platform**

Designed and implemented a pipeline to build, validate and release supplemental model for runtime augmentation of static global language model. This model is used for doing pronunciations hotfixes or for adjusting weights in the global model.

Amazon **May 2014 -Aug 2014**

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

Analyzed counterfeit spike problem for high volume items on Amazon's third party marketplace. I also designed a generic framework that flags and blocks the sale of dubious products based on a rule based criteria. The rules for the framework were derived from counterfeit spike analysis mentioned above.

Chatimity **Sept 2011-June 2013**

Software Engineer

First employee at Chatimity. Along with two founders, I helped built a scalable pseudo-anonymous chat based social network that handled millions of messages per day. Chatimity was recently acquired by Freshdesk.

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

System Software Engineer, Multimedia Audio Team

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

SELECTED PROJECTS:

Compositional Language Model (code)

Implemented the idea proposed in the paper "*Compositional Approach to Language Modeling.*" The framework is written in Java with DeepLearning4j framework and UJMP as sparse matrix library. Inside-Outside score calculation was done using the grammar from Stanford CoreNLP. The code generates both a language model and word embeddings with a compositional framework to embed sequences in the same space.

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 to four data sets, namely *RaceSpace*, *Digits*, *Faces* and *Swiss Roll*. The project was an individual effort and was done as a course project for Advanced Machine Learning class.

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

Comparative Evaluation of Supervised Learning Algorithms (code)

Did comparative analysis of Supervised Learning Algorithms in Python using Scikit-Learn and Theano. This project was done in a team of three for Machine Learning class. The data sets studied were: *Wisconsin Breast Cancer*, *Iris*, *Higgs*, *OCR,* and *Hand Writing Recognition.*

Algorithms evaluated were: *Multi-Layer Perceptron, 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 PIDGIN, an ontology alignment technique that uses web text as interlingua. In this project, we mapped ontologies for Freebase, NELL, and Yago to each other using label propagation algorithm. This project was done as an independent study in Data Science Lab under Dr. Daisy Wang and was a part of a 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

OTHERS:

Organizer for Deep Learning Reading Group for Alexa engineering audience.

REFERENCES:

[1] Strom, Nikko. "Scalable distributed dnn training using commodity gpu cloud computing." *INTERSPEECH*. Vol. 7. 2015.