Vivek Gupta

Ph.D Applicant in Machine Learning & Natural Language Processing

Contact Information

Microsoft Research India
#9, Vigyan, Lavelle Road
Bangalore, India - 560001

http://vgupta123.github.io
t-vigu@microsoft.com
+91-8765696260

Research Interests

Representation Learning, Extreme Classification, Text Summarization, Fairness & Diversity

Current Position

Microsoft Research India (June 2016 - Present)
Research Fellow Advisors: Dr. Prateek Jain & Dr. Nagarajan Natarajan

Education

Indian Institute of Technology Kanpur (July 2015 - May 2016)
M. Tech with Honors in Computer Science and Engineering, CGPA: 9.3/10.0

Indian Institute of Technology Kanpur (July 2011 - July 2015)
B. Tech with Honors in Computer Science and Engineering, CGPA: 7.5/10.0

Accepted Publications

Mekala, D.*,**Gupta, V.***,Paranjape, B., Karnick, H., *Sparse Composite Document Vectors using soft clustering over distributional representations*, **EMNLP 2017** (Oral) [**Paper**][**PPT**]

Gupta, V., Karnick, H.,Bansal, A., Jhala, P., *Product Classification in E-Commerce using Distributional Semantics* , **COLING 2016** (Poster) [**Paper**][**POS**]

Wadbude, R.,**Gupta, V.**, Mekala, D., Karnick, H., *User Bias Removal in Review Score Prediction*, Accepted in **CODS-COMAD 2018** (Long Oral) & **DAB 2017** (Workshop) [**Paper**][**POS**]

Gupta, V., Mittal, S.,Bhaumik, S., Roy, R. *Assisting Humans to Achieve Optimal Sleep by Changing Ambient Temperature*, **BIBM 2016**, **BHI 2016** & **HI-DS 2016** (Oral) [**Paper**][**PPT**]

Work Under Review

Wadbude, R., **Gupta, V.**, Rai, P., Natararjan, N.,Karnick, H., Jain, P., *Leveraging Distributional Semantics for Multi-Label Learning*, Under review in **ECML PKDD 2018** (Journal) **[PrePrint]**

Dohare, S., Karnick, H., **Gupta, V.**, *Text Summarization using Abstract Meaning Representation* Under review in **NAACL 2018** **[PrePrint]**

Mekala, Dheeraj., **Gupta, V.**, Kar, P., Karnick, H., *Bayes-optimal Hierarchical Classification over Asymmetric Tree-Distance Loss* Under preparation for **ICML 2018** **[PrePrint]**

Mahajan, D., **Gupta, V.**, Keerthi, S.,Sundararjan, S., *Efficient Estimation of Generalization Error and Bias-Variance Components of Ensembles*, Under review in **SDM 2018** **[PrePrint]**

Master Thesis

Product Categorization in E-Commerce using Distributional Semantics
Advisors: Prof. Harish Karnick (IIT Kanpur) & Pradhuman Jhala (Flipkart.com) [**Thesis**][**PPT**]

- Proposed a novel distributional semantics representation for text documents
- Proposed a two-level ensemble approach utilizing (with respect to the taxonomy tree) a path-wise, node-wise and depth-wise classifiers for product classification

Research Experience

ProtoSSL: Resource Constraint Anomaly Detection (June 2016 - Present)
Advisors: Dr. Prateek Jain(Microsoft Research India), Dr. Nagarajan Natarajan (Microsoft Research India) and Dr. Praneeth Netrapalli (Microsoft Research India)

- Working on a resource constrained machine learning algorithm for semi-supervised learning
- Working on predictive maintenance, investigating time series features, novel evaluation metrics for anomaly detection

Leveraging Distributional Semantics for Multi-Label Learning (June 2016 - May 2017)
Advisors: Dr. Nagarajan Natarajan (Microsoft Research India) Dr. Piyush Rai (IIT Kanpur), Dr. Harish Karnick (IIT Kanpur), Dr. Prateek Jain (Microsoft Research India)

- Design extreme multi-label learning algorithms using distributional semantics with more efficient training procedures with joint learning
- Extended the approach to naturally incorporate other sources of side-information, particularly the label-label co-occurrence matrix

Text Summarization using Abstract Meaning Representation (June 2016 - May 2017)
Advisor: Dr. Harish Karnick (IIT Kanpur)

- Explored a novel full-fledged pipeline for text summarization with an intermediate step of Abstract Meaning Representation (AMR)
- Proposed method achieves state-of the-art results compared to the other text summarization routines based on AMR
- Pointed out some significant problems in the existing evaluation methods making them unsuitable for evaluating summary quality

Text Categorization using Sparse Composite Document Vector (June 2016 - May 2017)
Advisor: Dr. Harish Karnick (IIT Kanpur)

- Proposed a novel topic-based document representation which outperforms state-of-the-art models in multi-class and multi-label classification tasks
- Showed that fuzzy GMM clustering on word-vectors lead to more coherent topic than LDA and can be used to detect Polysemic words
- Embeddings provide a robust estimation of the query and document language models, thus improving the MAP of language model based retrieval systems

Optimal Hierarchical Classification (June 2016 - Present)
Advisors: Prof. Purushottam Kar (IIT Kanpur) & Prof. Harish Karnick (IIT Kanpur)

- Extended the consistency of hierarchical classification algorithm on asymmetric tree distance loss using calibrated surrogates
- Showed under reasonable assumption over hierarchy that the Bayes optimal classification for this asymmetric loss can be found in $O(log(n))$ (Under preparation for ICML 2018)

Introducing Diversity Priors in Latent Variable Model (Jan 2016 - Apr 2016)
Advisor: Prof. Piyush Rai (IIT Kanpur), Report Link

- Worked on diversity priors regularizes for latent variable models like probabilistic PCA and non-negative matrix factorization
- Introduced log-det divergence, determinant point process and pair-wise Mutual angular regularizes in standard probabilistic models

Learning Bilingual Word Embeddings (Mar 2017 - Present)
Advisor: Dr. Raghavendra Udupa (Microsoft Research India)

- Working on learning transformations that project Word2Vec embeddings for a pair of languages onto a common embedding space
- Used word alignments obtained from IBM statistical MT models to train a model inspired by supervised semantic indexing (SSI) algorithm on samples of positive and negative word translations pairs

Learning Interpretable Word Embeddings (Mar 2017 - Present)
Advisor: Dr. Sundararajan Sellamanickam (Microsoft Research India)

- Working on learning interpretable word embedding by non negative and sparse factorization of PPMI matrix
- Evaluating embedding performance on Google Similarity and Analogy task and embedding interpretable on human intrusion task

Research Internship

Microsoft Research India, Bangalore (May 2016 - Jul 2016)
Estimation of generalization error for ensembles *Dr. Sundararajan Sellamanickam (Principal Applied Scientist)*

- Worked on efficient estimation of generalization error for ensembles using normality assumption on classification scores
- Worked on efficient prediction of ensemble parameters, bias and variance of generalization errors using minimal number of ensembles (Under Review SDM 2018)

Flipkart Internet Pvt. Ltd., Bangalore (May 2015 - July 2015)
Web Scale Product Classification *Pradhuman Jhala (Principal Architect)*

- Developed a model for deep hierarchal classification using multilevel classifiers and ensembles methods
- Model outperformed existing naive bayes classifier in precision@1, precision@1 and prediction time.

Samsung R&D Institute, Bangalore (May 2014 - July 2014)
Mobile and Healthcare Solution Y2014 *Sandip Bhaumik (Group Manager) & Raj Roy (Manager)*

- Did an innovation in S-Health Wearable technology on sleep applications by automatically changing ambient temperature using recurrent feedback signal from S-Watch (Patented & Published)

Scholastic Achievements

Selected among the top 7 finalist in **Ericson's Innovation Awards, India** in 2016
Selected in top 100 all over India in **Telnor, Internet for all challenge** by Unisys in 2015
Best poster presentation for Internship works at Flipkart Internet Private Ltd. in 2015
Best project award for project *Distributed Classroom* in Topic in Distributed System in 2015
Secured **All India Rank 183** in **IIT-JEE 2011** amongst more than 500,000 candidates
Qualified for **Kishore Vaigyanik Protsahan Yojana (KVPY)** scholarship in 2011
Selected among the top 1% of students in India, **Indian National Physics Olympiads** & **Indian National Chemistry Olympiads** in 2011
Selected among the top 1% of students in State, **Indian National Physics Olympiads** in 2010

Professional Service

External Reviewer: Served as a reviewer for long and short papers for Empirical Methods in Natural Language Processing 2017

Coordinator: Initiated and managed Special Interest Group in Machine Learning at Computer Science and Engineering Department, IIT Kanpur. Organize regular meetups for discussions and talks on topics in Machine Learning and related fields.

Teaching Assistant: for MLT 2016 - Machine Learning Tool and Technique: Mentored a group of 30 M-Tech students part of an introductory course on Machine Learning. Set up a labeling software for project work, resulting in a new dataset.

Student Secretary & **Student Mentor**: in Promotion of Work Experience and Research PoWER & Alumni Contact Program (ACA) under Office of Dean of Research and Development IIT Kanpur

Student Volunteer: for mentoring and teaching underprivileged students of primary classes from nearby village(Nankari) at Prayas, IIT Kanpur

Talks & Seminars

I had the opportunity to present my work at various places which I throughly enjoyed. [**Talks**]

Student Mentorship

I have been extremely lucky to have mentored some amazing students. [**Students Mentored**]

Travel Grants & Scholarship

I have been fortunate to receive scholarships/grants at several occasion to support my education and research. [**Travel Grants**]

Development Internship

Flipkart Internet Pvt. Ltd., Bangalore (Dec 2015 - Jan 2015)
Extracting Attribute Value Pairs *Pradhuman Jhala, Principal Architect*

- Developed a machine learned multilabel model for extraction of attribute values attribute values collar ,crew & vneck ,full sleeve & half sleeve for shirts {pointy, non-pointy} for sandals,{baseball, basketball, bracket & ballet} for sport shoes from product images with pair wise accuracy of 74%

Synopsys. Inc., Bangalore (May 2013 - July 2013)
Trend Generation for Dali Coverage *Yogesh Pandey, Director*

- Created interactive visualization of Scenario coverage trend for functional verification project *Dali* using data visualizations technique. Also, coded object oriented oilter pattern which parse json file and perform analysis on selected properties

Other Projects

Generalization vs Optimization for Deep Learning (June 2016 - Present)
Advisor: Dr. Praneeth Netrapalli, Microsoft Research India

- Working on more efficient techniques (tail-averaging) for non-convex optimization for training deep neural nets & investigated the broad question of role of neural nets depths help in generalization & optimization

Independence & Randomization in Deep Neural Net (Jan 2016 - Present)
Advisor: Prof. Harish Karnick(IIT Kanpur) & Gaurush Hiranandani (Adobe Research)

- Worked on information theoretic proof of *Improved Classification and Reconstruction using Independence and Randomization* published in DICTA 2015. Evaluated non-linear dimensionality reduction techniques by appropriately scoring topology preservation and information gain.

Large Scale Multi Kernel Learning (July 2015 - Dec 2015)
Advisor: Prof. Harish Karnick(IIT Kanpur) [**Report**][**PPT**]

- Showed that Multi Kernel Learning work better than single kernel independent data separability and distribution in space. Extended toolbox for multi-modal data using different kernels for data attributes from different heterogeneous sources.

Forest Cover Type Prediction (Dec 2014 - Apr 2015)
Advisor: Prof. Harish Karnick (IIT Kanpur) [**Report**][**PPT**]

- Predict forest type of forest from 7 classes using 54 attributes (kaggle data) and was ranked 42 with 82% accuracy on kaggle. Used classification algorithms like Random Forests, Extra-Trees Classifier, Auto-encoders, Gradient Boosting Model, K -Nearest Neighbor, Label propagation, Support Vector Machine etc.

Location Based Approximate Query (Dec 2014 - Apr 2015)
Advisor: Prof. Arnab Bhattacharya (IIT Kanpur) [**Report**]

- Modified a Filter Effective Hybrid tree (FEH) for indexing approx geo-location queries using filter tree based on distance metric like Euclidean & Manhattan. Result out perform in search time by significant margin for R* tree using line extra space. Developed a Graphical User Interface using Google Map API.

Distributed Classroom (PANGO) (Aug 2014 - Dec 2014)
Advisor: Prof. Ratan K Ghosh (IIT Kanpur) [**Report**][**PPT**][**Video**][**JED-I**]

- Developed a server-less distributed application for file sharing, digital whiteboard, conducting online tests, live lecture slidesharing, video and document tagging and automatic lecture scribing. Implemented various distributed protocols like token based leader election, mutual exclusion joining, deadlock prevention, novel distributed file sharing and non-buffering syncing protocol

Image Search Engine-PicBook (Aug 2014 - Dec 2014)
Advisor: Prof. TV. Prabhakar (IIT Kanpur) [**Report**][**PPT**][**Video**]

- Developed a collaborative social Image search engine (PICBOOK) using Neo4j graph database which capture collaborative clicks. Software support automating tagging, re-ranking, graph search queries like *Photos of my friend friends, my photos, female friend photos*, text search, uploads, friends addition