Statement of Purpose

Vivek Gupta User ID: keviv9@gmail.com

I am a first year Doctoral student at the School of Computing, University of Utah advised by Prof. Vivek Srikumar. My current research interest are Machine Learning and Natural Language Processing, particularly problems incorporating both fields. I am fortunate to work with Prof. Vivek Srikumar on problem of Semi-Structured Natural Language Inference. I am also exploring few topics in fair Machine Learning under the guidance of Prof. Suresh Venkatasubramanian. Before joining University of Utah, I was a *Research Fellow*[1] in the **Machine Learning and Natural Language Systems group** at **Microsoft Research India**. Previously, I completed my masters in Computer Science from IIT Kanpur, where I also started exploring my research interests in Machine Learning (ML) and Natural Language Processing (NLP) under guidance of Prof. Harish Karnick. I developed a keen interest in ML and NLP while working on product classification at Flipkart, an e-commerce company in India; in particular, leveraging techniques of NLP to solve ML problems and *vice versa.* I have applied ML techniques to solve complex NLP tasks (ML for NLP) like multi-document summarization, multi sentence document representation, and unbiased search recommendations. On the other side, I also used NLP techniques for solving complex ML problems (NLP for ML) like extreme multi label learning (XML), and resource constrained machine learning. Below, I elaborate on some of these problems along with the techniques used to solve them.

ML for NLP: For my master's thesis, I worked with Prof. Harish Karnick and Mr. Pradhuman Jhala on *Product Classification in E-Commerce using Distributional Semantics* [2]. Multiple tasks in e-commerce (e.g. search) require tagging of the textual description of a product with the path labels from a static hierarchical taxonomy. Such categorization is challenging because most categories have sparse and non-uniform number of products. It also requires good representation of product descriptions and an efficient algorithm for classification. To handle these challenges, I developed: (1) a novel document representation technique, and (2) an ensemble of multiple classifiers predicting path labels, node-wise labels and depth-wise labels. **This work was published in the International Conference on Computational Linguistics (COLING, 2016) (acceptance rate: 32%) [1]. Our framework was integrated into production at Flipkart, and was even covered by the media**[3]**. My master project was completely funded by Flipkart**. My project was directly supervised by top leadership namely Dr. Muthusamy Chelliah (Director, Academic Engagement, Flipkart).

Eager to explore more research problems in this area, I joined Microsoft Research India as a *Research Fellow* in the Machine Learning and Natural Language Systems group. At Microsoft, I have collaborated with researchers on several projects on ML and NLP. One interesting problem I worked was on representational learning. Multiple tasks in NLP like text categorization require good and efficient textual representation. Earlier techniques suffer from multiple challenges, viz - (1) documents are represented in same dimension as words; (2) semantic distinctiveness of words are ignored; (3) have large feature formation time; (4) multiple meanings of a word is ignored etc. To handle these challenges, I proposed a novel document representation technique called *Sparse Composite Document Vector (SCDV).* SCDV is sparse, has less feature formation time and outperforms previous state of art in multiple NLP tasks. **The work was published in EMNLP, 2017 [2]. I was delighted to know that the representation is being used by Bing Ads team at Microsoft to detect duplicate ads**. However, our representation require tuning of a sparsity parameter which increases deployment time. To overcome this, we are using dictionary learning with feature selection, which also improve our results, we submitted the extended work to **ICLR 2019 [3]**. We also Incorporated sense aware embedding with our representation. The work is currently under review at **NAACL-HLT 2019**.

NLP for ML: I have worked with Dr. Piyush Rai, Dr. Nagarajan Natarajan and Prof. Harish Karnick at Microsoft Research Lab, India on *Leveraging Distributional Semantics for Multi Label Learning.* In particular, our framework is on challenging extreme multi-label learning -: (1) Large scale setting (million labels, million learning examples and large feature dimension), (2) skewed and heavy tail distribution of labels, (3) missing labels in training and test dataset, and (4) requirement of diversity in prediction. My idea was to use distributional semantic algorithms for multi-label learning. We modified an existing state of the art extreme learning algorithm to improve the training time and cope with the problem of missing labels, while keeping test performance intact. Further, we also performed joint learning of embedding and regressors by using stochastic gradient descent on our novel objective. Our final performance is comparable and sometime better to the state of art algorithms for all standard datasets. Currently, we are exploring better text featurization techniques for multi label learning. **Our work is accepted as oral presentation at AAAI 2019 [4].**

Other ML Problems: I am also interested in the application of ML to solve problems of industrial importance. In particular, I worked on problems like predictive maintenance (anomaly detection), resource constrained machine learning, optimal hierarchical classification, efficient ensemble learning, cost-sensitive learning, and novel performance metrics. Some of these problems require understanding of aspects like fairness, interpretability and privacy, which ML systems should satisfy in addition to good performance. One problem which satisfy these industrial aspect I worked was on *Predictive Maintenance using Machine Learning.* Our objective was to develop an interpretable model which can

[1] The RF program is competitive program at Microsoft Research India, geared to prepare students for graduate studies [2] Thesis: `https://vgupta123.github.io/thesis.html` [3] Financial Express `https://goo.gl/Fcfyij` & Economic Times `https://goo.gl/xSV6QZ`

detect machine downtime while keeping false alarms within limited budget. I developed multiple classification models on real world data to predict machine downtimes using past sensor data (time-series). Working with industrial data is challenging: a) temporal nature of data (time-series), b) redundant and noisy data, and c) requirement to predict ahead. **Our model was demonstrated to top leadership and customers in Microsoft, who finally deployed it into live production**. Currently, we are looking at sequence to sequence based models for anomaly detection. Apart from the above problem, I am exploring the problem of *Resource Constrained Semi-Supervised Learning.* In semi-supervised setting, only a small subset of training examples is provided along with significantly large number of unlabeled examples. The objective is to develop a k-nearest neighbors algorithm learning models, with fewer and sparse candidate points, in each class. We are modifying an existing label propagation algorithm for resource constraint setting i.e. lower memory footprint and faster prediction. . I have worked on theoretical problems (provable machine learning) as well, e.g. (1) Efficient Estimation of Generalization Error [5] with Dr. Sundararajan Sellamanickam, and (2) Bayes Optimal Hierarchal Classification [6] with Prof. Purushottam Kar.

The challenges associated with such problems and potential impact created by solving them, motivated me to pursue a PhD. I wish to continue working on similar challenging problems in future. By working at several places [4], particularly at Microsoft Research & IIT Kanpur, I have developed a principled and nuanced approach to research. In this process, I got exposed to complete research life cycle, from problem ideation to real-world application. My long-term career objective is to be a leading contributor in academic and industrial-research, and to mentor others like me. **I have initiated and managed a Special Interest Group in Machine Learning (SIGML) [5] for promoting ML/NLP interests at IIT Kanpur**. Eminent researchers from academia and industry give seminars talks in the group. SIGML has organized reading group sessions, machine learning research days and has motivated many IIT Kanpur students to pursue ML research. **I worked as a teaching assistant for the course Machine Learning, Tools and Techniques** at IIT Kanpur. **I have also mentored few students in their undergraduate/master project [6], some of which resulted in publications [7]**. I had the opportunity to present my work at various places [8], which I throughly enjoyed. I have maintained a good academic track record throughout my academia with current 3.7/4 GPA score at University of Utah and 9.3/10 GPA in masters at IIT Kanpur. I am looking for opportunity where I can make contribution to research with significant practical impacts. After brief discussion with Prof. Vivek Srikumar, I wish to work either on Commonsense Reasoning or Information Extraction projects under guidance of Jonathan May and Violet (Nanyun) Peng at the USC/ISI. I choose the above two projects because of common research interest, familiarity with problems and my confidence in my ability to perform well. However, I am also open to other interesting and challenging problems. I believe that my research experience, personal initiative and passion towards research give my candidature a strong impetus.

References

[1] Vivek Gupta, Harish Karnick, Ashendra Bansal, and Pradhuman Jhala. Product classification in e-commerce using distributional semantics. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 536–546. The COLING 2016 Organizing Committee, 2016.

[2] Dheeraj Mekala*, Vivek Gupta*, Bhargavi Paranjape, and Harish Karnick. Scdv : Sparse composite document vectors using soft clustering over distributional representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680. Association for Computational Linguistics, 2017.

[3] Vivek Gupta, Ankit Kumar Saw, Partha Pratim Talukdar, and Praneeth Netrapalli. Unsupervised document representation using partition word-vectors averaging, 2019.

[4] Rahul Wadbude*, Vivek Gupta*, Piyush Rai, Nagarajan Natarajan, Harish Karnick, and Prateek Jain. Leveraging distributional semantics for multi-label learning. *CoRR (Under review ECML-PKDD 2018)*, abs/1709.05976, 2017.

[5] Dhruv Mahajan, Vivek Gupta, S Sathiya Keerthi, Sellamanickam Sundararajan, Shravan Narayanamurthy, and Rahul Kidambi. Efficient estimation of generalization error and bias-variance components of ensembles. *CoRR*, abs/1711.05482 (Under review SDM 2018), 2016.

[6] Dheeraj Meka, Vivek Gupta, Purushottam Kar, and Harish Karnick. Bayes-optimal hierarchical classification over asymmetric tree-distance loss. *Under preparation for ICML 2018 (https://goo.gl/hBzPHf)*, 2018.

[7] Vivek Gupta*, Siddhant Mittal*, Sandip Bhaumik, and Raj Roy. Assisting humans to achieve optimal sleep by changing ambient temperature. In *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on Bioinformatics and Biomedicine*, pages 841–845. IEEE, 2016.

[8] Rahul Wadbude, Vivek Gupta, Dheeraj Mekala, and Harish Karnick. User bias removal in fine grained sentiment analysis. *CoRR (Accepted at CoDS-COMAD 2018 & DAB@CIKM 2017)*, abs/1612.06821, 2016.

[9] Shibhansh Dohare, Harish Karnick, and Vivek Gupta. Text summarization using abstract meaning representation. *CoRR (Submitted to NAACL 2018)*, abs/1706.01678, 2017.

[4] Synopsys Inc, Samsung Research [7], Flipkart.com, IIT Kanpur, Microsoft Research [5] `https://www.cse.iitk.ac.in/users/sigml/`
[6] Student Mentored: `https://vgupta123.github.io/mentor.html` [7] User Bias Removal: [8] (CoDS-CoMAD, 2018 & DAB workshop CIKM, 2017), Text Summarization: [9](Preprint) [8] `https://vgupta123.github.io/talks.html`