Distributional Semantics meets Multi Label Learning

Vivek Gupta[^#], Rahul Wadbude*, Nagarajan Natarajan[#]
Harish Karnick*, Prateek Jain[#] and Piyush Rai*

[^]School of Computing, University of Utah
[#] Microsoft Research Lab, India
*Indian Institute of Technology, Kanpur

The Thirty-Third AAAI Conference on Artificial Intelligence,
AAAI 2019

*** Rahul is now @ (AlphaGrep, Bangalore)**

work done when Vivek was at Microsoft

Outline

- Problem Statement
- Challenges
- Proposed Technique
- Experimental Results
- Conclusion
- Takeaway Points

Classification Paradigms

Pick one

Label 1	✓
Label 2	

Binary

Pick one

Label 1	
Label 2	
Label 3	
Label 4	✓
...	
...	
Label L	

Multi-class

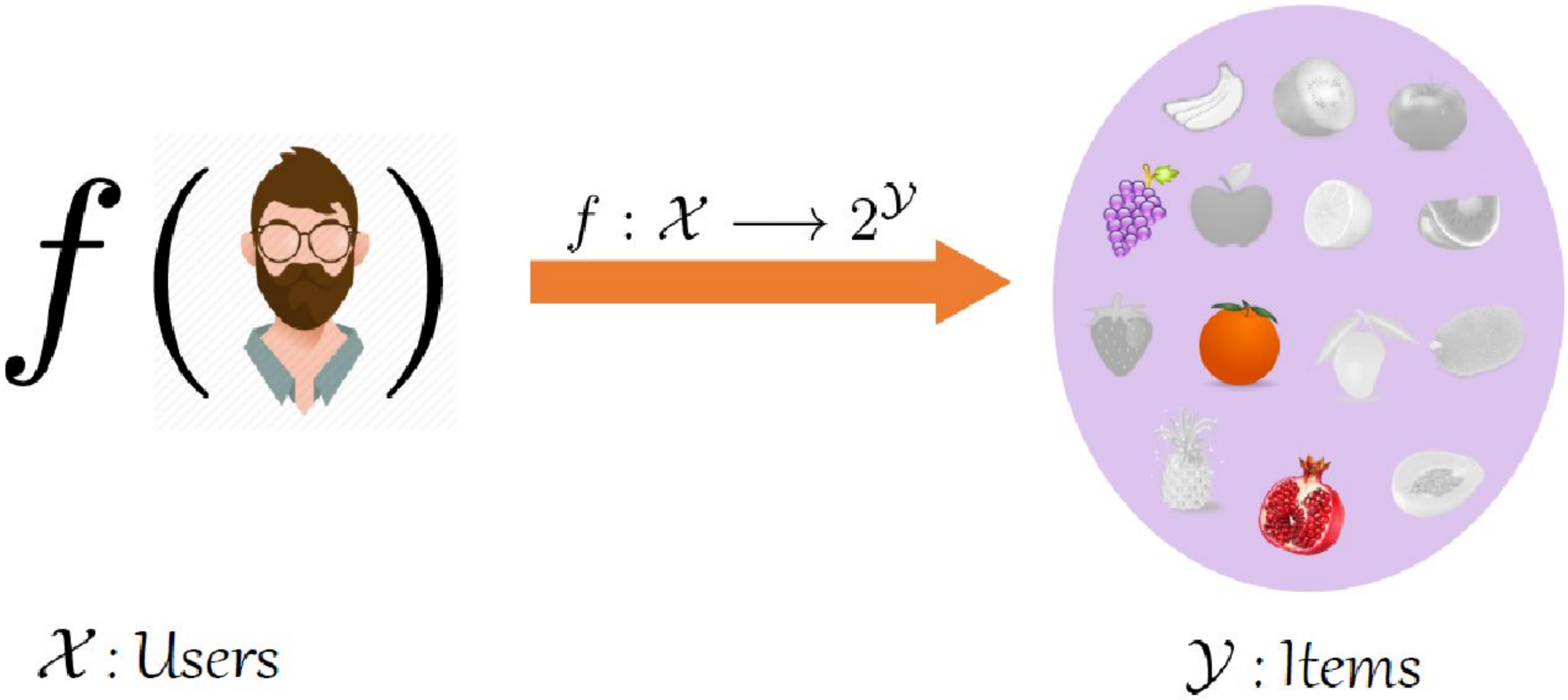
Pick all applicable

Label 1	
Label 2	✓
Label 3	
Label 4	✓
...	
...	
Label L	✓

Multi-label

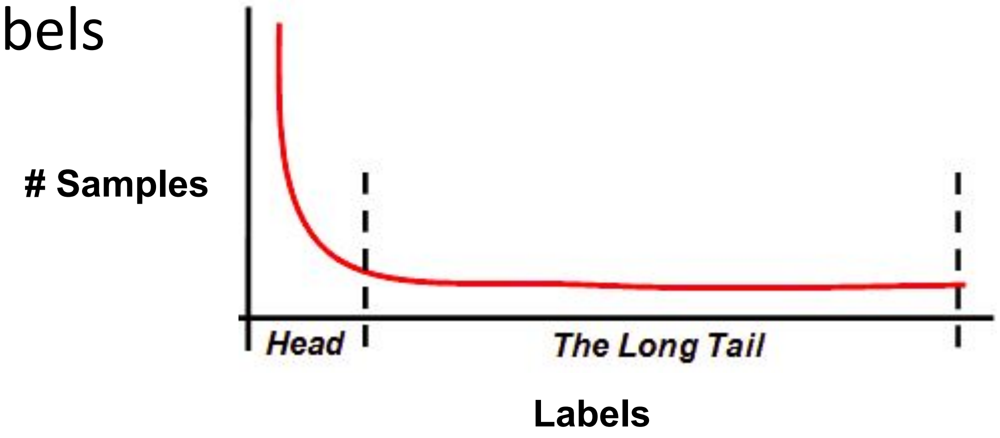
Extreme Multi-Label Learning

What all items would this user buy?



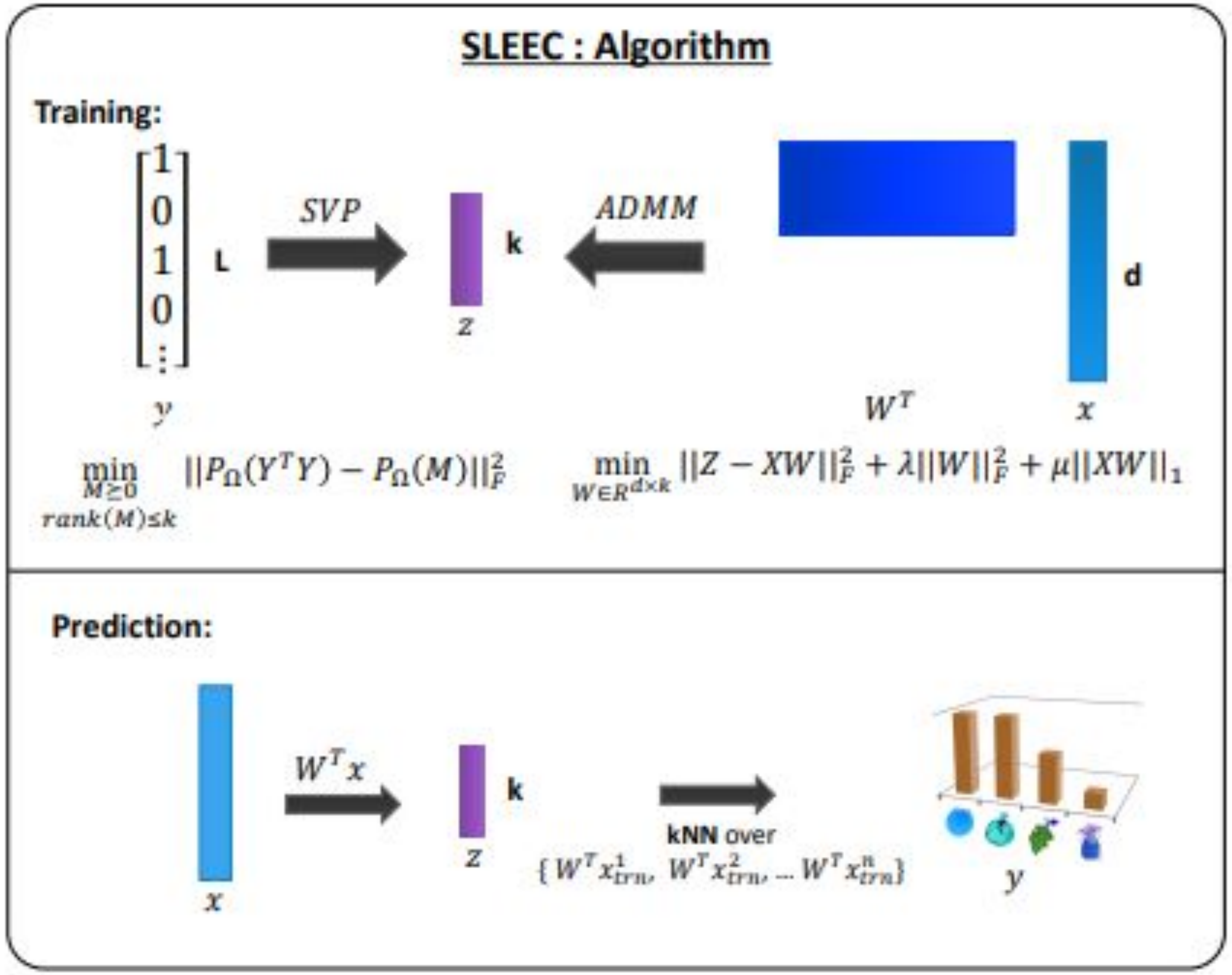
Challenges and Opportunity

- Large scale setting
    - N (#examples), L (#labels), D (#Feature Dim) in millions
    - Challenging due to long tail distribution of Labels
- Missing label in training and prediction set
    - Exploiting label correlation
    - Appropriate training and evaluation



SLEEC - Embedding Based Algorithm

Non linear neighborhood preserving low rank embedding of label vectors



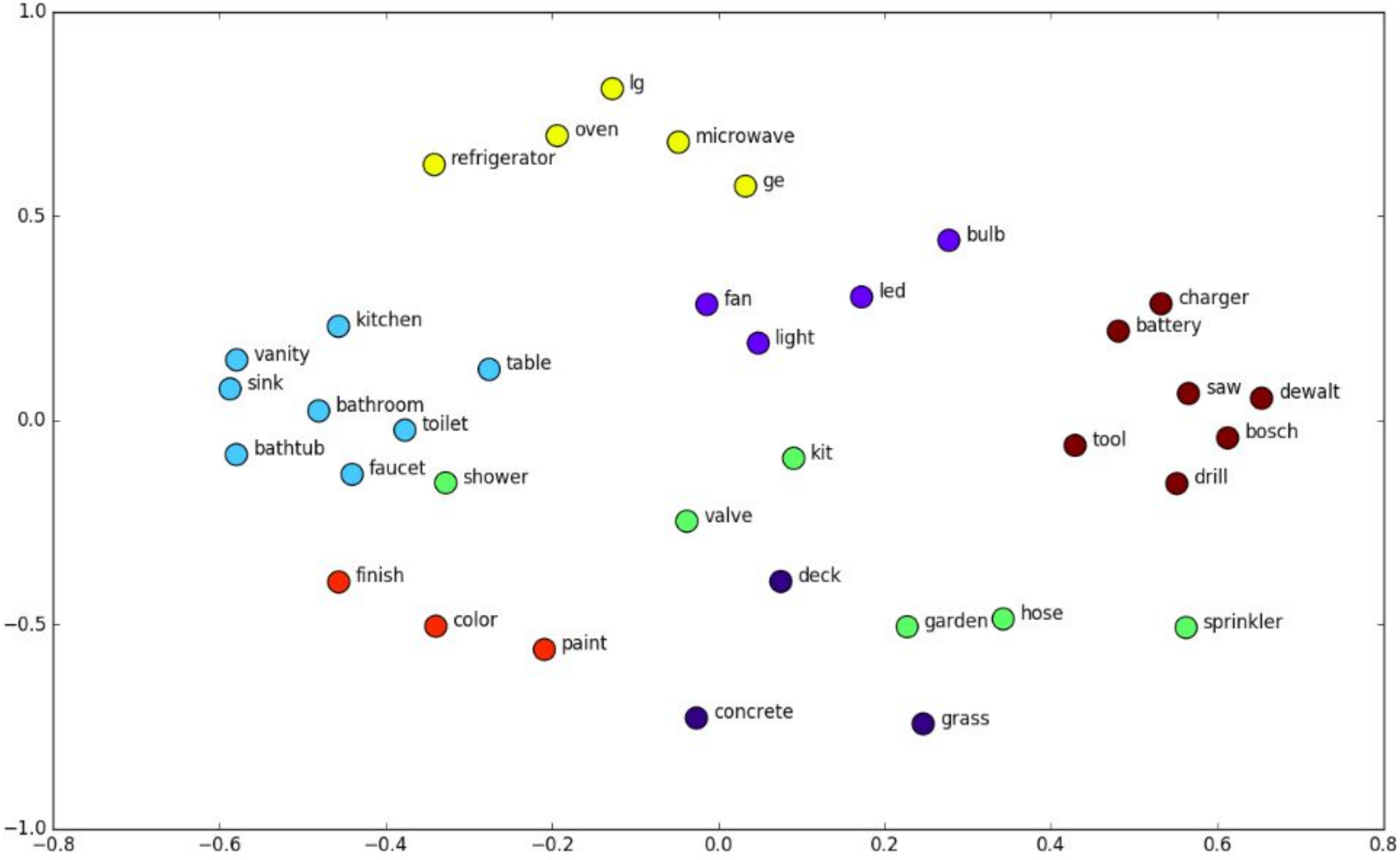
- Inefficient in training time
- Cannot perform end-to-end joint learning
- Cannot handle missing label

SLEEC : Kush Bhatiya and Himanshu Jain. *"Sparse Local Embeddings for Extreme Multi-label Classification"* , in *NeurIPS*, 2015.

Contribution

- Novel objective that **leverages the *word2vec* embedding** methods
- Optimized **efficiently by matrix factorization**, thus faster than *SLEEC*
- Can do **joint learning of embedding-regression**, more accurate than SLEEC
- Can **easily incorporates side information**, thus handling the missing labels

word2vec



Similar words are found in similar locations (src: http://suriyadeepan.github.io)

SGNS meets Label Embedding

- **word2vec** embedding using **Skip Gram Negative Sampling** objective

$$P(\text{Observing } (w, w')) = \sigma(\langle \mathbf{z}_w, \mathbf{z}_{w'} \rangle) = \frac{1}{1 + \exp(\langle -\mathbf{z}_w, \mathbf{z}_{w'} \rangle)}$$

$$\max_{\mathbf{z}} \sum_{w} \left(\sum_{w':(w',w)} \log\left(\sigma(\langle \mathbf{z}_w, \mathbf{z}_{w'} \rangle)\right) + \frac{n_-}{\#w} \sum_{w''} \log\left(\sigma(-\langle \mathbf{z}_w, \mathbf{z}_{w''} \rangle)\right) \right)$$

- **replacing words** with the **instance label vectors** in the training sets

$$\max_{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n} \sum_{i=1}^{n} \left(\sum_{j:\mathcal{N}_k(\mathbf{y}_i)} \log\left(\sigma(\langle \mathbf{z}_i, \mathbf{z}_j \rangle)\right) + \frac{n_-}{n} \sum_{j'} \log\left(\sigma(-\langle \mathbf{z}_i, \mathbf{z}_{j'} \rangle)\right) \right)$$

SGNS as Matrix Factorization

Theorem

SGNSs objective is equivalent to weighted matrix factorization of shifted positive point wise mutual information (SPPMI) matrix [8]

Shifted PPMI:

$$PMI_{ij}(M) = log\left(\frac{M_{ij} * |M|}{\sum_k M_{(i,k)} * \sum_k M_{(k,j)}}\right)$$

$$SPPMI_{ij}(M) = \max(PMI_{ij}(M) - log(k), 0)$$

Here, PMI is point wise mutual information matrix of M and $|M|$ represent sum of all element in M.

Proposed ExMLDS Algorithm

- Overall, **multi-iteration SVP replaced** with **single step SVD**
- Regression and prediction algorithm remain same as in *SLEEC*.
- We observe the ***ExMLDS*** **training is 10x faster** than *SLEEC*.

Incorporating Label-Label Correlation

- Learn the **embeddings of labels as well as instances jointly**.
- **Overall Idea** :
  - think of **labels as individual words**
  - think of **instances with the active label as sentence**
- Use **extra label correlation information** for label embedding
- Helps in **handling the missing label problem efficiently**

SLEEC Joint Learning

- Joint learning objective for the *SLEEC* algorithm

$$\min_{V \in \mathbb{R}^{\widehat{L} \times d}} \|P_\Omega(Y^T Y) - P_\Omega(X^T V^T V X)\|_F^2 + \lambda \|V\|_F^2 + \mu \|V X\|_1.$$

- it's highly non-convex as well as non-differentiable

ExMLDS Jointly Learning

- With our proposed objective?

$$\mathbb{O}_i = \sum_{j:\mathcal{N}_k(\mathbf{y}_i)} \log\left(\sigma(K_{ij})\right) + \frac{n_-}{n} \sum_{j'} \log\left(\sigma(-K_{ij'})\right),$$

- Joint learning possible, although non-convex nature

$$\nabla_V \mathbb{O}_i = \sum_{j:\mathcal{N}_k(\mathbf{y}_i)} \sigma(-K_{ij}) \nabla_V K_{ij} - \frac{n_-}{n} \sum_{j'} \sigma(K_{ij'}) \nabla_V K_{ij'}$$

$$\nabla_V K_{ij} = -ab^3 c \mathbf{z}_i (\mathbf{x}_i)^T - abc^3 \mathbf{z}_j (\mathbf{x}_j)^T + bc(\mathbf{z}_i \mathbf{x}_j^T + \mathbf{z}_j \mathbf{x}_i^T)$$

$$a = \mathbf{z}_i^T \mathbf{z}_j, b = \frac{1}{||\mathbf{z}_i||}, c = \frac{1}{||\mathbf{z}_j||}$$

Efficient Training

Method	Bibtex	Delicious	Eurlex	Mediamill	Delicious-200K
ExMLDS1	**23**	**259**	**580.9**	**1200**	**1937**
ExMLDS2	143.19	781.94	880.64	12000	13000
SLEEC	313	1351	4660	8912	10000

Training time

Dataset	Prec@k	Proposed ExMLDS1	Embedding Based DXML	SLEEC	LEML
Bibtex	P@1	63.38	63.69	**65.29**	62.54
	P@3	38.00	37.63	**39.60**	38.41
	P@5	27.64	27.71	**28.63**	28.21
Delicious	P@1	67.94	67.57	**68.10**	65.67
	P@3	61.35	61.15	61.78	60.55
	P@5	56.3	56.7	57.34	56.08
Eurlex	P@1	77.55	77.13	79.52	63.40
	P@3	64.18	64.21	64.27	50.35
	P@5	52.51	52.31	52.32	41.28
Mediamill	P@1	87.49	**88.71**	87.37	84.01
	P@3	**72.62**	71.65	72.6	67.20
	P@5	**58.46**	56.81	58.39	52.80
Delicious-200K	P@1	46.07	44.13	**47.50**	40.73
	P@3	41.15	39.88	**42.00**	37.71
	P@5	38.57	37.20	**39.20**	35.84

Performance

ExMLDS1 much faster than SLEEC with almost equal performance

Performance with Missing Labels

Dataset	Prec@k	ExMLDS3	Sleec	Leml	Leml-Imc
Bibtex	P@1	**48.51**	30.5	35.98	41.23
	P@3	**28.43**	14.9	21.02	25.25
	P@5	**20.7**	9.81	15.50	18.56
Eurlex	P@1	**60.28**	51.4	26.22	39.24
	P@3	**44.87**	37.64	22.94	32.66
	P@5	**35.31**	29.62	19.02	26.54
rcv1v2	P@1	**81.67**	41.8	64.83	73.68
	P@3	**52.82**	17.48	42.56	48.56
	P@5	**37.74**	10.63	31.68	34.82

We hide randomly **80%** of the labels from training labels. We provide extra **YY' (original)** complete label-label correlation matrix along with masked **Y** to both *LEML-IMC* and *ExMLDS3*.

Performance with Joint Learning

Dataset	Prec@k	Proposed EXMLDS4	Embedding Based ANNEXML	SLEEC	XML-CNN
AmazonCat-13K	P@1	93.05	93.55	90.53	**95.06**
	P@3	79.18	78.38	76.33	**79.86**
	P@5	64.54	63.32	61.52	**63.91**
Wiki10K-31K	P@1	**86.82**	86.50	85.88	84.06
	P@3	**74.30**	74.28	72.98	73.96
	P@5	63.68	**64.19**	62.70	64.11
Delicious-200K	P@1	47.70	46.66	**47.85**	-
	P@3	41.22	40.79	**42.21**	-
	P@5	37.98	37.64	**39.43**	-
WikiLSHTC-325K	P@1	62.15	63.36	54.83	-
	P@3	39.58	40.66	33.42	-
	P@5	29.10	29.79	23.85	-
Wikipedia-500K	P@1	62.27	**63.86**	58.39	59.85
	P@3	41.43	**42.69**	37.88	39.28
	P@5	31.42	**32.37**	28.21	29.31
Amazon-670K	P@1	41.47	42.08	35.05	-
	P@3	36.35	36.65	31.25	-
	P@5	32.43	32.76	28.56	-

Conclusions

- **Novel objective for XML** that **leverages the *word2vec* embedding** method
- **Optimized efficiently by matrix factorization**, making it's **faster than *SLEEC***
- Objective **can jointly learn** and **obtain better results** compared to *SLEEC*
- **Easily incorporates side information**, that is **useful for handling missing labels**

Takeaway Point

- **Distributional Semantics** algorithms can be efficiently utilize for XML task

- **Joint learning of embedding and regression** could be beneficial for XML task

Questions to Ponder?

- Can we **jointly embed instance feature (*x*) and instance label (y)** for XML task ?

- Better method for **selection of negative samples** while instance embedding ?

Paper id: #6621

I am active seeking for summer 2019 research internship opportunity.
In case of available suitable position, please let me know:

keviv9@gmail.com
https://vgupta123.github.io

Acknowledgement

- Anonymous reviewers whose reviews help in improving the paper quality
- AAAI-19 Student Scholar and Volunteer Program for the needful support
- Prof. Vivek Srikumar and Prof. Ellen Riloff of School of Computing, University of Utah for presentation review at a short notice
- Microsoft Research Lab,Bangalore; School of Computing, University of Utah and Indian Institute of Technology, Kanpur for needed support and guidance

Main References

- **ExMLDS** : Rahul Wadbude and Vivek Gupta. *“Leveraging Distributional Semantics for Multi Label Leaning”*. AAAI 2019
- ***Word2vec*** *:* Tomas Mikolov and Ilya Sutskever. *“Distributed representations of words and phrases and their compositionality"* In: NeurIPS 2013
- **Doc2vec** : Quoc V Le and Tomas Mikolov. “Distributed Representations of Sentences and Documents” In ICML 2014
- **SLEEC** : Kush Bhatiya and Himanshu Jain. *“Sparse Local Embeddings for Extreme Multi-label Classification”* , in *NeurIPS,* 2015.
- **FastXML** : Yashoteja Prabhu, and Manik Varma. “*A Fast, Accurate and Stable Tree-classifier for eXtreme Multi-label Learning”*, in *KDD,* 2014
- **PfastreXML** : Himanshu Jian and Yashoteja Prabhu. “*Extreme Multi-label Loss Functions for Recommendation, Tagging, Ranking & Other Missing Label Applications*”, in *KDD*, 2016.
- **PDSparse** : I. E. H. Yen and X. Huang. “*PD-Sparse: A Primal and Dual Sparse Approach to Extreme Multiclass and Multilabel Classification”* in *ICML*, 2016.
- **DiSMEC** : R. Babbar, and B. Schölkopf. “*DiSMEC - Distributed Sparse Machines for Extreme Multi-label Classification”*, in *WSDM* 2017.
- **PPDSparse** : I. E. H. Yen and X. Huang. “PPDSparse: A Parallel Primal-Dual Sparse Method for Extreme Classification” in *KDD*, 2017.
- **AnnexML** : Yukihiro Tagami. *“Approximate Nearest Neighbor Search for Extreme Multi-label Classification”, in KDD 2017*
- **Slice:** H. Jain and V. Balasubramanian, **“***Scalable linear extreme classifiers trained on 100 million labels for related searches”, in WSDM 2019*
- **Parabel:** Y. Prabhu and A. Kag**, “***Partitioned label trees for extreme classification with application to dynamic search advertising”, in WWW 2018*
- **SwiftML:** Y. Prabhu and A. Kag, *“Extreme multi-label learning with label features for warm-start tagging, ranking and recommendation”, in WSDM 2018*