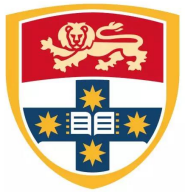
Deep Streaming Label Learning

Zhen (Zohn) Wang, Liu Liu, Dacheng Tao

UBTECH Sydney AI Centre
School of Computer Science



THE UNIVERSITY OF
SYDNEY

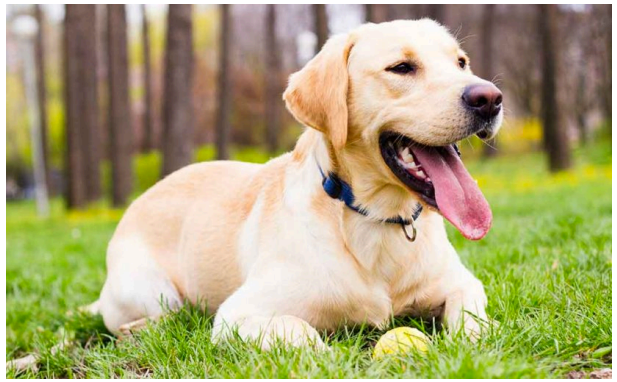
Outline

- Background and motivation
- Proposed approach
- Theoretical analysis
- Experimental analysis
- Conclusion and future work

Background and Motivation

Traditional supervised learning (Multi-class/Single-label learning)

labels: [‘dog’, ‘person‘, ‘cat’]



label vectors: [1,0,0] [0,1,0] [0,0,1]

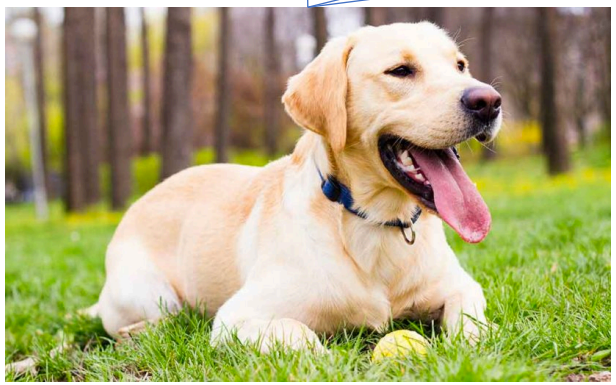
Multi-label learning

labels:
‘dog’,
‘person‘,
‘cat’

[1,1,0]

[1,1,1]

[1,0,1]



[1,0,0]

[0,1,0]

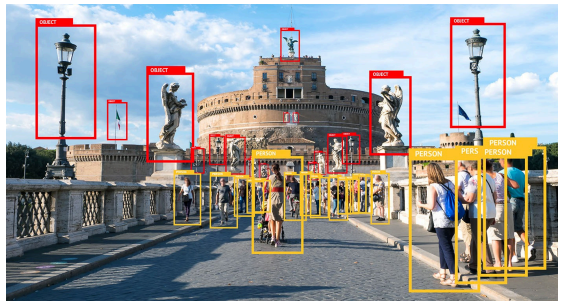
[0,0,1]

Streaming label learning

past labels:
‘dog’,
‘person‘,
‘cat’

emerging new labels:
‘grass’,
‘shirt‘,
‘smile’
…

Social network:
tag photos or texts continuously



Detect system:
emerge novel events

Interest groups:
form new interests and groups

Streaming Label Learning: **model newly arrived labels expediently and effectively**

Streaming label learning

Goal and challenge:

How?

Effectiveness: building an effective new-label classifier (leveraging historical knowledge)

Efficiency: without retraining the past-label classifier (only training on new labels)

Our previous work [1] presented a simple solution by using a linear classifier, which trains a learnable matrix for new labels with the relationship between past labels and new labels.

$$J(\boldsymbol{w}_{m+1}) = \sum_{i=1}^{n} \ell(y^*_{m+1,i}, \mathbf{x}_i^T \boldsymbol{w}_{m+1}) + \frac{\beta}{2} \|\boldsymbol{w}_{m+1} - W_m \mathbf{s}_{m+1}\|_2^2$$

Weaknesses of [1]:

- linear representation limits performance;
- training knowledge from classifiers of past labels is neglected.

[1] You, S., Xu, C., Wang, Y., Xu, C., and Tao D., "Streaming label learning for modeling labels on the fly," arXiv, 2016.

Proposed Approach

Deep Streaming Label Learning (DSLL)

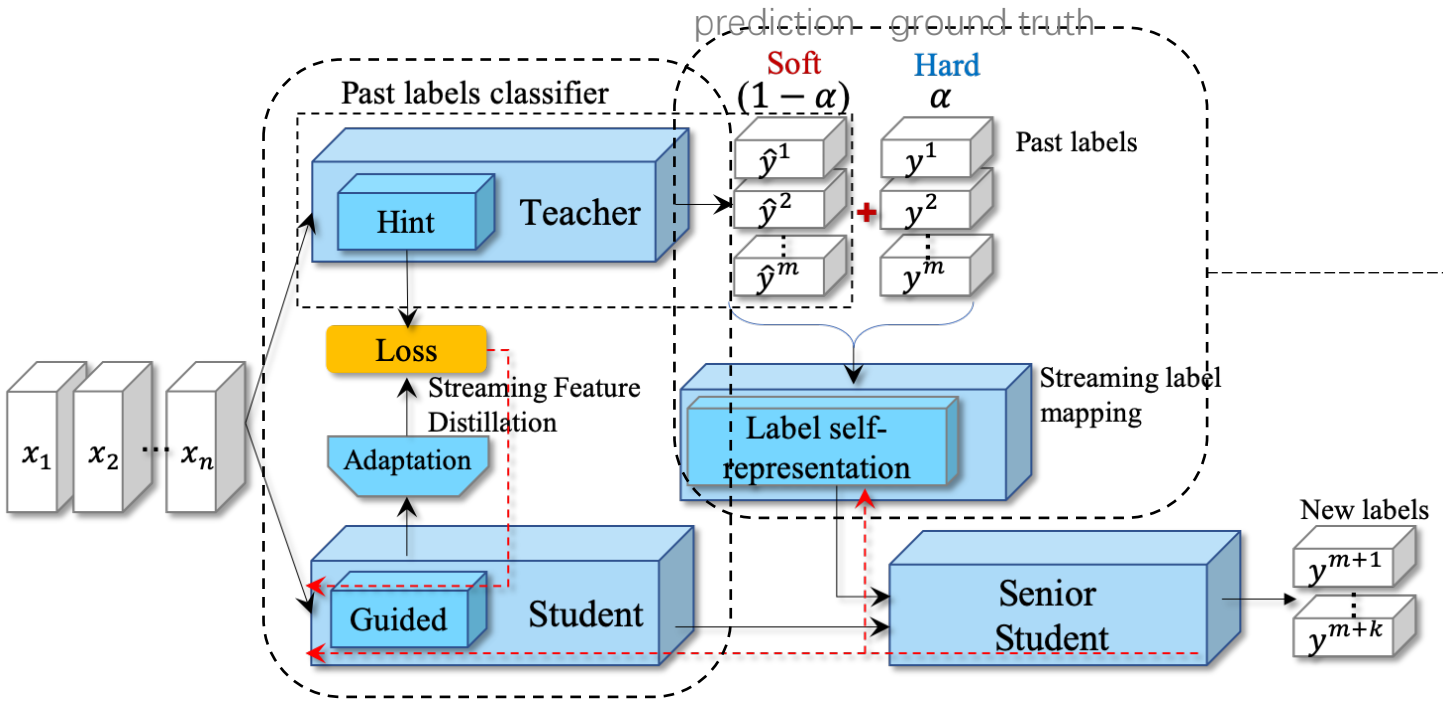
Deep streaming label learning

In this paper, we propose a DNN-based framework, Deep Streaming Label Learning (DSLL), to classify instances with newly emerged labels.

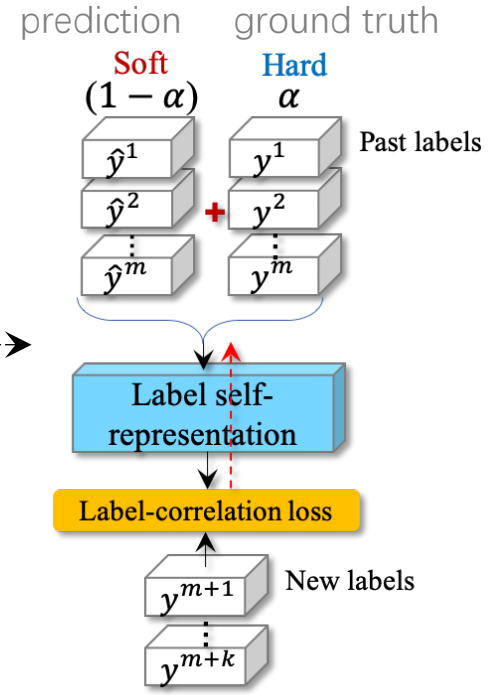
DSLL can

- explore and utilize deep correlations between past labels and new labels;
- leverage the knowledge from previous learning to model new labels without retraining the whole multi-label classifier.

Deep streaming label learning (DSLL)



The overall DSLL architecture



(1) Streaming Label Mapping

✓ captures the relationship between new labels and past labels with a proposed label correlation-aware loss.

$$\mathcal{L}_{\mathbf{S}}(\tilde{\boldsymbol{y}},\boldsymbol{y}^{new})=$$
$$\sum_{i=1}^{n}\frac{1}{|Y_i^+||Y_i^-|}\sum_{(p,q)\in Y_i^+\times Y_i^-}\|(\mathbf{S}_m(\tilde{\boldsymbol{y}}_i)^p-\mathbf{S}_m(\tilde{\boldsymbol{y}}_i)^q)-b\|_2^2$$

(2) Streaming Feature Distillation

✓ transforms the knowledge from the teacher to the new-label classifier (student).

(3) Senior Student Network

✓ leverages the relationship and knowledge to finally model new labels. The red dotted lines denote the backpropagation path during learning.

Theoretical Analysis

Theoretical analysis

We theoretically analyze the excess risk bounds for our learning model and provide a generalization error bound for the new-label classifier DSLL.

DSLL learns a classifier $\widehat{\mathbf{W}}$ by minimizing the empirical risk,

$$\hat{\mathbf{W}}=\underset{\mathbf{W}\in\mathcal{W}}{\operatorname{argmin}}\hat{\mathcal{L}}(\mathbf{W})=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=m+1}^{m+k}\ell(y_i^j,f_i^j(\boldsymbol{x}_i^*,\mathbf{W})),$$

Our goal is to show that the learned classifier $\widehat{\mathbf{W}}$ is generalizable,

$$\mathcal{L}(\hat{\mathbf{W}})\leq\inf_{\mathbf{W}\in\mathcal{W}}\mathcal{L}(\mathbf{W})+\epsilon,$$

where population risk of a classifier, defined as:

$$\mathcal{L}(\mathbf{W}):=\mathbb{E}_{(\boldsymbol{x}^*,\boldsymbol{y}^{new})}\left[\left[\sum_{j=m+1}^{m+k}\ell(\boldsymbol{y}^j,f^j(\boldsymbol{x}^*,\mathbf{W}))\right]\right].$$

Theoretical analysis

We theoretically analyze the excess risk bounds for our learning model and provide a generalization error bound for the new-label classifier DSLL.

Theorem 1. Suppose we learn a new classifier $\mathbf{W}$ in terms of k new labels using formulation $\widehat{\mathbf{W}} = \arg\min_{\mathbf{W}\in\mathcal{W}} \hat{\mathcal{L}}(\mathbf{W})$ over a set of n training instances. Then, with a probability of at least $1-\delta$, we have

$$\mathcal{L}(\hat{\mathbf{W}}) \leq \inf_{\mathbf{W}\in\mathcal{W}} \mathcal{L}(\mathbf{W}) + \mathcal{O}\left(\sqrt{\frac{k}{n}}\left(\gamma_x + \gamma_y\sqrt{k}\right)\right) + \mathcal{O}\left(k\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right),$$

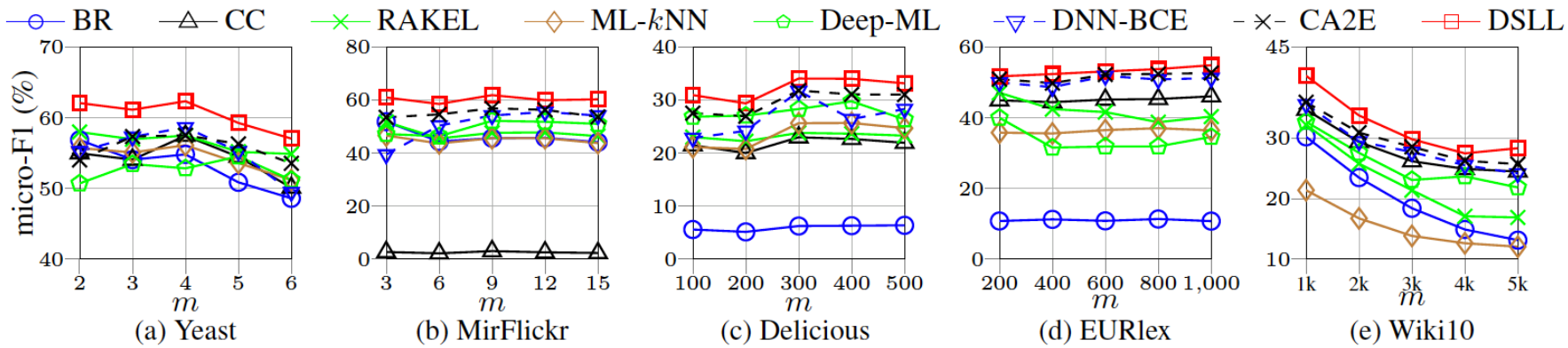
where γ_x and γ_y are real constant numbers related to the number of neurons in the output layer. We assume, without loss of generality, that $\mathbb{E}[\![\|\boldsymbol{x}\|_2^2]\!] \leq 1, \mathbb{E}[\![\|\boldsymbol{y}^{new}\|^2]\!] \leq k$.

Experimental Analysis

Experimental analysis

5 multi-label datasets
10 compared methods
10 evaluation metrics

Dataset	Domain	#Training	#Testing	#Feature	#Labels	#Card-Features	#Card-Labels
yeast	biology	1,500	917	103	14	103.00	4.24
MirFlickr	image	10,417	2,083	1,000	38	539.33	4.74
Delicious	web	12,920	3,185	500	983	18.17	19.03
EURlex	text	15,479	3,869	5,000	3,993	5.31	236.69
Wiki10	text	14,146	6,616	101,938	30,938	673.45	18.64



Performance comparison of learning new labels with different batch sizes by considering 50% of labels as past labels. m indicates the number of new labels.

Experimental analysis

Ranking performance of each comparison algorithm for learning new labels with different batch sizes by regarding 50% of labels as past labels.

- ✓ The results show that DSLL outperforms the other approaches with respect to handling newly arrived labels.

Datasets	#label	micro-AUC ↑										
		BR	CC	RAKEL	ML-*k*NN	SLEEC	SML	SLL	Deep-ML	DNN-BCE	C2AE	DSLL
yeast	2	0.6703	0.6750	0.6914	0.6731	0.7003	0.7114	0.7209	0.7228	0.7616	0.7624	**0.7793**
	3	0.7140	0.6972	0.7162	0.7074	0.8003	0.8287	0.7302	0.8437	0.8521	0.8590	**0.8641**
	4	0.6978	0.7096	0.7113	0.7027	0.7751	0.7955	0.7401	0.8216	0.8383	0.8384	**0.8530**
	5	0.6883	0.6979	0.7061	0.6964	0.7775	0.7889	0.7203	0.8139	0.8325	0.8318	**0.8507**
	6	0.6768	0.6743	0.7059	0.6834	0.7645	0.7751	0.7133	0.8115	0.8079	0.8199	**0.8229**
MirFlickr	3	0.6589	0.5049	0.6572	0.6296	0.5476	0.6123	0.7291	0.7448	0.7592	0.8051	**0.8122**
	6	0.7100	0.5045	0.7187	0.6624	0.6400	0.6959	0.8388	0.8476	0.8606	0.8628	**0.8713**
	9	0.7311	0.5071	0.7170	0.6754	0.7001	0.7331	0.8671	0.8692	0.8778	0.8890	**0.8972**
	12	0.7151	0.5057	0.7223	0.6692	0.6862	0.7116	0.8630	0.8600	0.8763	0.8807	**0.8886**
	15	0.7157	0.5051	0.7253	0.6611	0.6805	0.7015	0.8624	0.8584	0.8787	0.8749	**0.8873**
Delicious	100	0.7133	0.5658	0.6382	0.5707	0.7813	0.8274	0.7981	0.8776	0.8615	0.8545	**0.8820**
	200	0.7080	0.5606	0.6400	0.5692	0.7815	0.8258	0.7956	0.8626	0.8677	0.8321	**0.8888**
	300	0.7177	0.5719	0.6493	0.5925	0.7941	0.8387	0.8068	0.8792	0.8655	0.8777	**0.9037**
	400	0.7159	0.5705	0.6515	0.5916	0.7964	0.8141	0.8059	0.8736	0.8909	0.8656	**0.9048**
	500	0.7144	0.5679	0.6445	0.5868	0.7926	0.8124	0.8030	0.8762	0.8697	0.8657	**0.9011**
EURlex	200	0.6936	0.7308	0.7015	0.6255	0.8591	0.8216	0.8813	0.8130	0.8201	0.8561	**0.8884**
	400	0.6738	0.7294	0.6821	0.6228	0.8481	0.8611	0.8905	0.8257	0.8423	0.8633	**0.8928**
	600	0.6769	0.7321	0.6743	0.6256	0.8547	0.8735	0.8995	0.8218	0.8472	0.8414	**0.9001**
	800	0.6825	0.7349	0.6575	0.6283	0.8548	0.8775	0.9016	0.8289	0.8491	0.8637	**0.9034**
	1000	0.6733	0.7361	0.6708	0.6258	0.8406	0.8691	**0.9115**	0.8246	0.8431	0.8456	0.9106
Wiki10	1k	0.6293	0.6535	0.6403	0.5694	0.8092	0.8113	0.8345	0.6978	0.7830	0.8274	**0.8406**
	2k	0.5928	0.6244	0.6029	0.5510	0.7505	0.7557	0.7876	0.6545	0.6968	0.7937	**0.8013**
	3k	0.5703	0.6078	0.5815	0.5405	0.7567	0.7192	0.7417	0.6505	0.7197	0.7939	**0.8023**
	4k	0.5567	0.6008	0.5687	0.5364	0.7198	0.7105	0.7366	0.6556	0.7307	0.7944	**0.7996**
	5k	0.5502	0.5984	0.5651	0.5345	0.7087	0.7194	0.7350	0.6463	0.7228	0.7824	**0.7865**

Datasets	#label	Ranking loss ↓										
		BR	CC	RAKEL	ML-*k*NN	SLEEC	SML	SLL	Deep-ML	DNN-BCE	C2AE	DSLL
yeast	2	0.3064	0.2912	0.2519	0.2966	0.1527	0.1493	0.1668	0.1538	0.1494	0.1483	**0.1439**
	3	0.3059	0.3217	0.2863	0.3064	0.0927	0.0901	0.1674	0.0840	0.0818	0.0812	**0.0807**
	4	0.3409	0.3722	0.3535	0.3433	0.1137	0.1093	0.1891	0.1111	0.1077	0.1114	**0.1065**
	5	0.3694	0.4113	0.3787	0.3731	0.1245	0.1179	0.2189	0.1230	0.1163	0.1242	**0.1147**
	6	0.4080	0.4606	0.4049	0.4136	0.1558	0.1508	0.2382	0.1492	0.1536	0.1514	**0.1475**
MirFlickr	3	0.3032	0.5336	0.2842	0.3157	0.2393	0.2253	0.1179	0.1051	0.1025	0.0881	**0.0809**
	6	0.2868	0.5786	0.2735	0.3337	0.2488	0.2189	0.0586	0.0571	0.0565	0.0569	**0.0562**
	9	0.3366	0.6854	0.3194	0.3981	0.2186	0.2386	0.0672	0.0663	0.0609	0.0580	**0.0570**
	12	0.4011	0.7816	0.3784	0.4690	0.2692	0.2528	0.0877	0.0865	0.0857	0.0772	**0.0750**
	15	0.4209	0.7966	0.3873	0.4929	0.2857	0.2710	0.0899	0.0909	0.0837	0.0858	**0.0812**
Delicious	100	0.4135	0.6551	0.4550	0.6521	0.2333	0.2235	0.1438	0.0916	0.0890	0.0929	**0.0830**
	200	0.5115	0.8147	0.5603	0.7988	0.2712	0.2523	0.1771	0.1168	0.1217	0.1224	**0.1140**
	300	0.5248	0.8303	0.5710	0.7918	0.2759	0.2594	0.1758	0.1093	0.1071	0.1073	**0.1031**
	400	0.5311	0.8360	0.5711	0.7978	0.2661	0.2608	0.1778	0.1102	0.1079	**0.1050**	0.1082
	500	0.5383	0.8435	0.5841	0.8096	0.2784	0.2679	0.1822	0.1109	0.1154	0.1094	**0.1067**
EURlex	200	0.1473	0.1281	0.1403	0.1787	0.0654	0.0743	0.0198	0.0879	0.0337	0.0205	**0.0138**
	400	0.2493	0.2048	0.2386	0.2880	0.1101	0.0901	0.0299	0.1062	0.0380	0.0274	**0.0226**
	600	0.3468	0.2898	0.3494	0.4041	0.1552	0.1325	0.0404	0.1848	0.0562	0.0531	**0.0313**
	800	0.4072	0.3400	0.4384	0.4780	0.1871	0.1618	0.0483	0.1530	0.0609	0.0495	**0.0330**
	1000	0.4861	0.3960	0.4939	0.5619	0.2375	0.2085	0.0585	0.2326	0.0754	0.0721	**0.0370**
Wiki10	1k	0.4746	0.4375	0.4213	0.5520	0.0918	0.1082	0.0867	0.2259	0.1011	0.0483	**0.0421**
	2k	0.6205	0.5642	0.5705	0.6818	0.2385	0.2052	0.1259	0.3373	0.2225	0.0648	**0.0606**
	3k	0.7324	0.6577	0.6823	0.7766	0.3306	0.2573	0.1897	0.3727	0.1911	0.0715	**0.0642**
	4k	0.8167	0.7246	0.7569	0.8424	0.4037	0.3029	0.2123	0.3899	0.1831	0.0751	**0.0687**
	5k	0.8538	0.7495	0.7941	0.8698	0.4313	0.4368	0.2234	0.4525	0.1972	0.0833	**0.0816**

Experimental analysis

Ablation Study

	Baseline	KD	CE	LC	LC-$\mathbf{S}_t$	$\mathbf{L}_{hard}$	$\mathbf{L}_{soft}$	$\mathbf{L}_{soft}^{hard}$	KD+LC-$\mathbf{S}_t$+$\mathbf{L}_{soft}^{hard}$
Average precision ↑	0.3824	0.4018	0.4113	0.4182	0.4287	0.4159	0.4427	0.4592	0.4925
Coverage ↓	0.2415	0.2349	0.2245	0.2221	0.2188	0.2352	0.2289	0.2237	0.2153

- ✓ The results show that DSLL outperforms the other approaches with respect to handling newly arrived labels.
- ✓ We evaluate different strategies for learning knowledge from the past and confirm the effectiveness of various parts of our proposed framework.

Conclusion and future work

We propose Deep Streaming Label Learning (DSLL) to address the multi-label learning problem with emerging new labels.

Properties

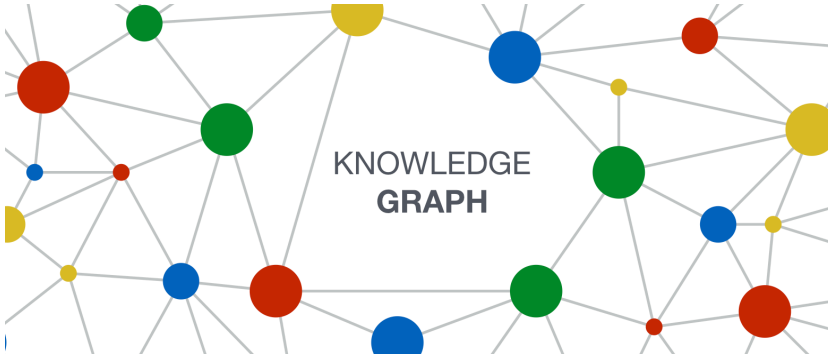
- ✓ Exploiting historical knowledge to model new labels effectively,
- ✓ Without computationally intensive

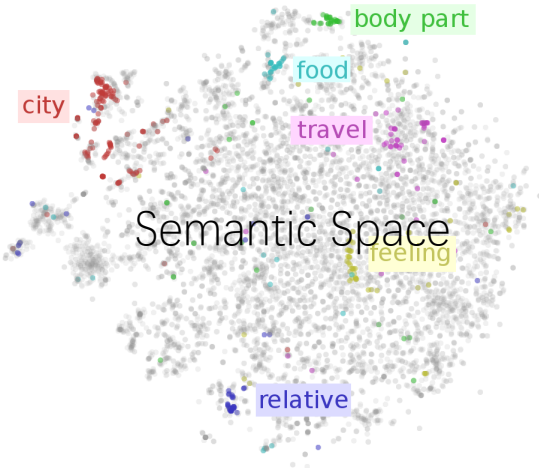
Components

- ✓ Streaming label mapping,
- ✓ Streaming feature distillation,
- ✓ Senior student.

Analyses

- ✓ Theoretical analysis prove a tight excess risk bound.
- ✓ Experimental analysis shows that DSLL achieves state-of-the-art performance and demonstrates its effectiveness.





Few-shot Learning Task

Thank you for your time 😄

THE UNIVERSITY OF SYDNEY

*all pictures are from the Internet.