LREC-COLING 2024

Self-Knowledge Distillation for Knowledge Graph Embedding

Presenter: Haotian Xu
Authors: Haotian Xu Yuhua Wang Jiahui Fan

PART 1

INTRODUCTION

Introduction

The general KGE method will obtain better performance by increasing the dimension of embedding.
Not only the number of model parameters but also the cost of training time will greatly increase with the rise of the embedding dimension.

Knowledge distillation is proposed to apply to KGE.
However, existing offline and online distillation methods usually cannot avoid the problem of computational sources and run-in memory consumed by the complex teacher model.

To avoid the above problems, we plan to apply self-knowledge distillation (SKD) to KGE.
SKD uses the distillation from the latest batch to generate soft targets to guide the training in the current batch.
We use dynamic temperature distillation (DTD) to design dynamic sample-wise temperatures to compute soft targets.
Knowledge adjustment (KA) is used to fix the predictions of misjudged training samples.

PART 2

METHODS

Self-Knowledge Distillation

$$L_{BCE}(\boldsymbol{y}_i, \boldsymbol{p}_i) = -\frac{1}{K}\sum_{k=1}^{K}\left(y_i(k)\log p_i(k) + \left(1 - y_i(k)\right)\log\left(1 - p_i(k)\right)\right)$$

$$L_{Hard} = \frac{1}{n}\sum_{i=1}^{n} L_{BCE}\left(\boldsymbol{y}_i, \boldsymbol{p}_i^{j}\right) \#$$

$$L_{KD} = \frac{1}{n}\sum_{i=1}^{n} \tau^2 D_{KL}(\tilde{\boldsymbol{p}}_i^{\tau} || \boldsymbol{p}_i^{\tau})$$

$$L_{Soft} = \frac{1}{n}\sum_{i=1}^{n} \tau^2 D_{KL}(\boldsymbol{p}_i^{\tau,j-1} || \boldsymbol{p}_i^{\tau,j})$$

Dynamic Temperature Distillation

The distillation temperature is too low → The model rarely focus on results that are much lower than the average.

The distillation temperature is too high → The model will confuse some similar categories with large probability.

A sample that easily confuses the model → Set the distillation temperature low

Easy to learn sample → Set the distillation temperature high

$$L_{DTD} = \frac{1}{n}\sum_{i=1}^{n} \tau_d^2 D_{KL}(\boldsymbol{p}_i^{\tau_d, j-1} || \boldsymbol{p}_i^{\tau_d, j})$$

$$\tau_d = \tau_0 + \left(\frac{\sum_{i=1}^{n} \omega_i}{n} - \omega_d \right) \beta$$

$$\omega_d = \frac{1}{s_{max}}$$

Dynamic Temperature Distillation

A confusing sample → Increase ω_d → The distillation temperature τ_d decrease

Easy to learn sample → Decrease ω_d → The distillation temperature τ_d increase

The maximum of logical vector s_{max} represents the model's confidence in the sample

The less confidence in sample → The more chaotic the sample → ω_d bigger

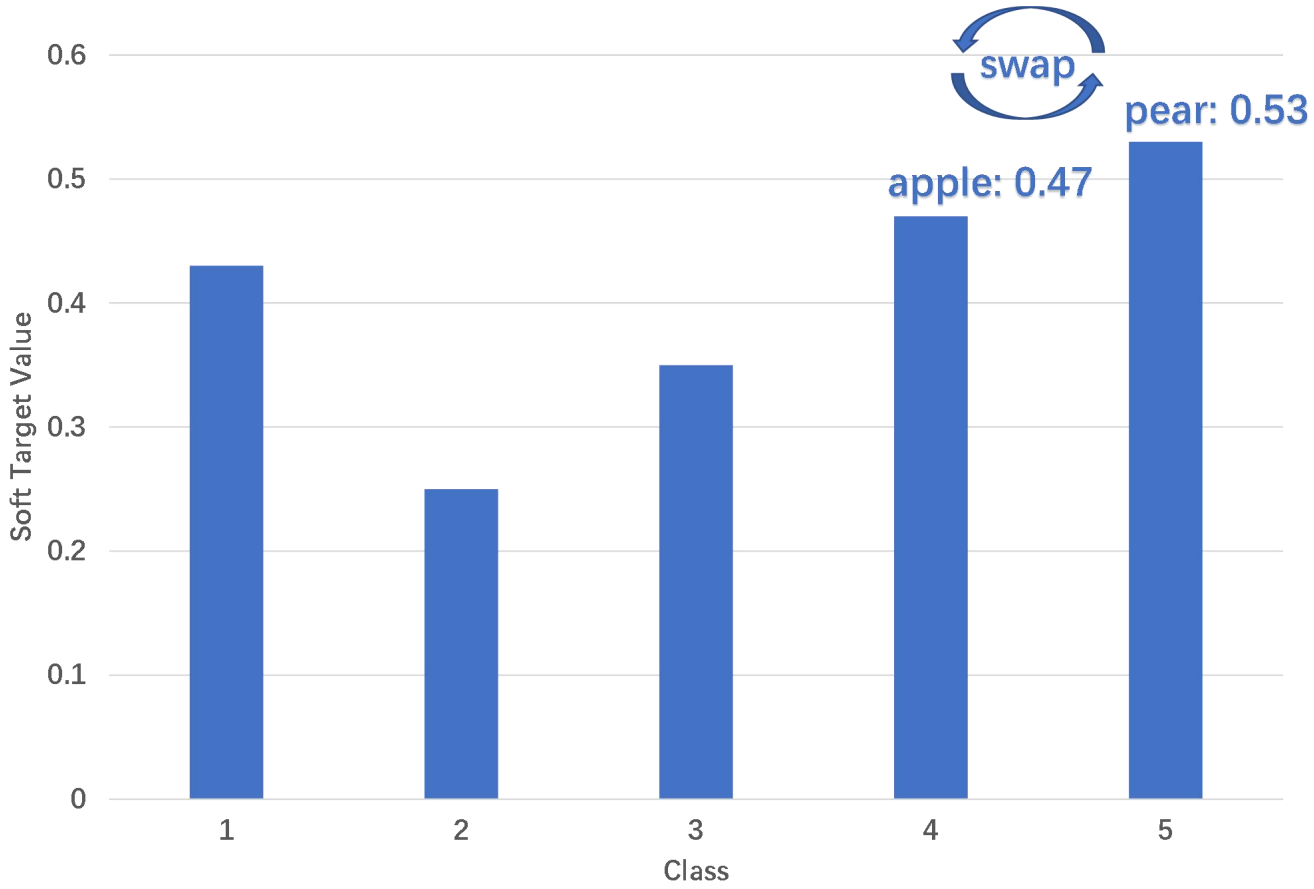
The more confidence in sample → The easier the sample is to learn → ω_d smaller

$$L_{DTD} = \frac{1}{n}\sum_{i=1}^{n} \tau_d^2 D_{KL}(\boldsymbol{p}_i^{\tau_d, j-1} || \boldsymbol{p}_i^{\tau_d, j})$$

$$\tau_d = \tau_0 + \left(\frac{\sum_{i=1}^{n} \omega_i}{n} - \omega_d \right) \beta$$

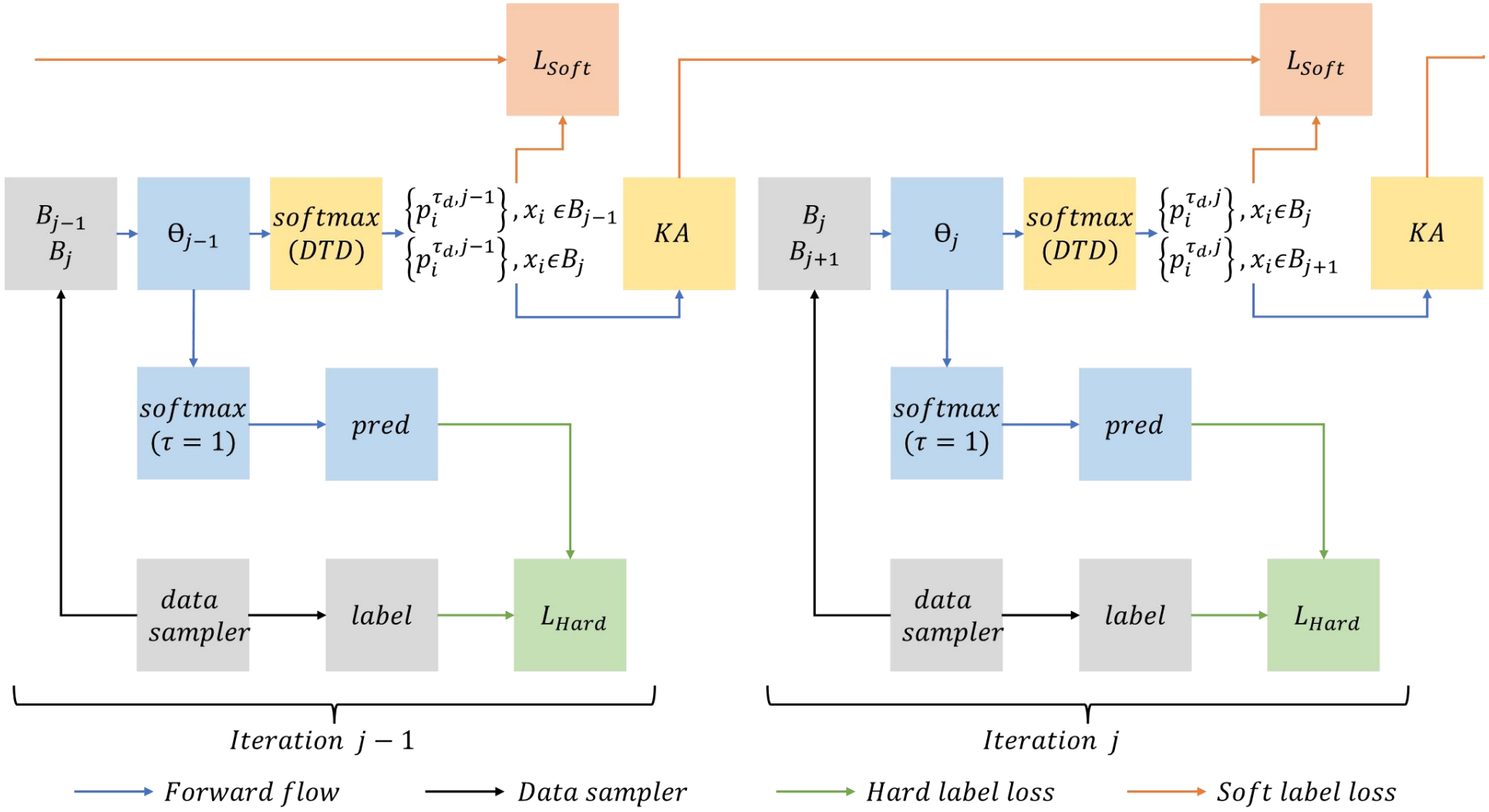
$$\omega_d = \frac{1}{s_{max}}$$

Knowledge Adjustment



$$L_{KA} = \frac{1}{n}\sum_{i=1}^{n} \tau_d^2 D_{KL}(f_a(\boldsymbol{p}_i^{\tau_d, j-1}) || \boldsymbol{p}_i^{\tau_d, j})$$

The SKDE module



$$L = L_{Hard} + \alpha \frac{1}{n} \sum_{i=1}^{n} \tau_d^2 D_{KL}(f_a(\boldsymbol{p}_i^{\tau_d,j-1}) || \boldsymbol{p}_i^{\tau_d,j})$$

PART 3

EXPERIMENTS

Experimental result

In the same dimension, SKDE's model performance improved significantly compared to the baseline model under all evaluation indicators.

These results can prove the effectiveness and generalization ability of our SKDE.

Model	WN18RR				FB15k-237			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
DistMult [1]	0.43	0.39	0.44	0.49	0.241	0.155	0.263	0.419
SKDE	**0.45**	**0.41**	**0.46**	**0.52**	**0.339**	**0.248**	**0.373**	**0.523**
SKDE (LW)	0.44	0.40	0.45	0.51	0.335	0.245	0.367	0.516
ComplEx [2]	0.44	0.41	0.46	0.51	0.247	0.158	0.275	0.428
SKDE	**0.47**	**0.44**	**0.49**	**0.54**	**0.349**	**0.257**	**0.383**	**0.534**
SKDE (LW)	0.46	0.43	0.48	0.53	0.339	0.249	0.371	0.522
ConvE [3]	0.43	0.40	0.44	0.50	0.311	0.223	0.339	0.493
SKDE	**0.44**	**0.41**	**0.46**	**0.53**	**0.326**	**0.236**	**0.356**	**0.508**
SKDE (LW)	**0.44**	0.40	0.45	0.52	0.324	0.234	0.355	0.506
AcrE(S) [4]	0.44	0.40	0.45	0.51	0.324	0.244	0.363	0.481
SKDE	**0.47**	**0.43**	**0.48**	**0.55**	**0.341**	**0.249**	**0.375**	**0.525**
SKDE (LW)	0.43	0.38	0.46	0.52	0.334	0.245	0.367	0.511
AcrE(P) [4]	0.45	0.42	0.46	0.52	0.328	0.247	0.367	0.485
SKDE	**0.48**	**0.44**	**0.49**	**0.55**	**0.353**	**0.260**	**0.390**	**0.540**
SKDE (LW)	0.41	0.35	0.44	0.51	0.339	0.250	0.372	0.517

The performance of SKDE after lightweight

Adding SKD, DTD and KA to the KGE does not change the number of parameters.

Reduced the dimension of entities and relations from 200 to 100.

The number of model parameters decreases with the decrease of vectorization dimension.

Model	WN18RR	FB15k-237
DistMult [1]	8193800	3003800
SKDE	8193800	3003800
SKDE (LW)	4096900	1501900
ComplEx [2]	16387600	6007600
SKDE	16387600	6007600
SKDE (LW)	8193800	3003800
ConvE [3]	10181299	4964897
SKDE	10181299	4964897
SKDE (LW)	4599299	1977897
AcrE(S) [4]	10813329	5596927
SKDE	10813329	5596927
SKDE (LW)	4796529	2175127
AcrE(P) [4]	11435873	6219471
SKDE	11435873	6219471
SKDE (LW)	4938873	2317471

The performance of SKDE after lightweight

After the dimension of SKDE is reduced, the performance of lightweight model (SKDE_LW) is slightly decreased.

The performance of SKDE_LW can still outperform the baseline model.

Model	WN18RR				FB15k-237			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
DistMult [1]	0.43	0.39	0.44	0.49	0.241	0.155	0.263	0.419
SKDE	**0.45**	**0.41**	**0.46**	**0.52**	**0.339**	**0.248**	**0.373**	**0.523**
SKDE (LW)	0.44	0.40	0.45	0.51	0.335	0.245	0.367	0.516
ComplEx [2]	0.44	0.41	0.46	0.51	0.247	0.158	0.275	0.428
SKDE	**0.47**	**0.44**	**0.49**	**0.54**	**0.349**	**0.257**	**0.383**	**0.534**
SKDE (LW)	0.46	0.43	0.48	0.53	0.339	0.249	0.371	0.522
ConvE [3]	0.43	0.40	0.44	0.50	0.311	0.223	0.339	0.493
SKDE	**0.44**	**0.41**	**0.46**	**0.53**	**0.326**	**0.236**	**0.356**	**0.508**
SKDE (LW)	**0.44**	0.40	0.45	0.52	0.324	0.234	0.355	0.506
AcrE(S) [4]	0.44	0.40	0.45	0.51	0.324	0.244	0.363	0.481
SKDE	**0.47**	**0.43**	**0.48**	**0.55**	**0.341**	**0.249**	**0.375**	**0.525**
SKDE (LW)	0.43	0.38	0.46	0.52	0.334	0.245	0.367	0.511
AcrE(P) [4]	0.45	0.42	0.46	0.52	0.328	0.247	0.367	0.485
SKDE	**0.48**	**0.44**	**0.49**	**0.55**	**0.353**	**0.260**	**0.390**	**0.540**
SKDE (LW)	0.41	0.35	0.44	0.51	0.339	0.250	0.372	0.517

Ablation Study

(1) After removing DTD and KA, the performance of the model is still better than that of the baseline model, indicating that SKD alone can optimize the model performance.

(2) DTD or KA alone can improve the performance of the model, which shows that both can optimize the performance of the model.

(3) Removing DTD results in a significant performance drop, indicating that DTD play a more important role.

(4) Models perform best when using both DTD and KA.

Model	WN18RR				FB15k-237			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
SKDE (DistMult)	**0.446**	**0.408**	**0.463**	**0.521**	0.3396	0.2483	**0.3735**	**0.5234**
-DTD	0.442	0.407	0.458	0.512	0.3378	0.2475	0.3695	0.5199
-KA	0.443	0.407	0.457	0.516	0.3372	0.2472	0.3694	0.5198
-KA and DTD	0.442	**0.408**	0.456	0.507	**0.3402**	**0.2507**	0.3727	0.5193
SKDE (ComplEx)	**0.472**	**0.439**	**0.486**	**0.536**	**0.3493**	**0.2572**	**0.3839**	**0.5341**
-DTD	0.468	0.435	0.482	0.528	0.3443	0.2528	0.3781	0.5285
-KA	0.470	0.438	**0.486**	0.531	0.3474	0.2568	0.3811	0.5298
-KA and DTD	0.469	0.437	0.484	0.530	0.3452	0.2538	0.3789	0.5281
SKDE (ConvE)	**0.442**	**0.406**	**0.455**	**0.527**	**0.3258**	**0.2360**	**0.3563**	**0.5081**
-DTD	0.438	0.400	0.448	0.521	0.3211	0.2312	0.3511	0.5040
-KA	0.437	0.399	0.449	0.523	0.3234	0.2348	0.3519	0.5033
-KA and DTD	0.436	0.398	0.447	0.520	0.3217	0.2315	0.3537	0.5032
SKDE (AcrE(S))	**0.467**	**0.427**	**0.481**	**0.552**	**0.3412**	**0.2493**	**0.3754**	**0.5246**
-DTD	0.463	0.423	0.477	0.545	0.3356	0.2445	0.3694	0.5191
-KA	0.457	0.419	0.470	0.539	0.3368	0.2465	0.3692	0.5207
-KA and DTD	0.455	0.417	0.466	0.537	0.3364	0.2484	0.3718	0.5189
SKDE (AcrE(P))	**0.477**	**0.440**	**0.488**	**0.549**	**0.3529**	**0.2595**	**0.3900**	**0.5397**
-DTD	0.470	0.436	0.480	0.540	0.3492	0.2581	0.3826	0.5343
-KA	0.471	0.433	0.484	0.542	0.3486	0.2558	0.3830	0.5371
-KA and DTD	0.469	0.432	0.483	0.540	0.3490	0.2572	0.3880	0.5326

PART 4

CONCLUSION AND FUTURE WORK

Conclusion

We propose a method to apply SKD, KA and DTD to KGE, called SKDE.

Extensive experiments demonstrate the effectiveness and generalization ability of our SKDE.

We achieve a lightweight model while maintaining a good model performance.

THANKS