中国科学技术大学
University of Science and Technology of China

美团

DisenKGAT: Knowledge Graph Embedding with Disentangled Graph Attention Network

Authors: Junkang Wu, Wentao Shi, Xuezhi Cao, Jiawei Chen*, Wenqiang Lei, Fuzheng Zhang, Wei Wu, Xiangnan He

Paper: https://arxiv.org/abs/2108.09628

Lab: USTC Lab for Data Science

Code: https://github.com/Wjk666/DisenKGAT

Date: Sep 29, 2021

CIKM 2021
1-5 NOVEMBER

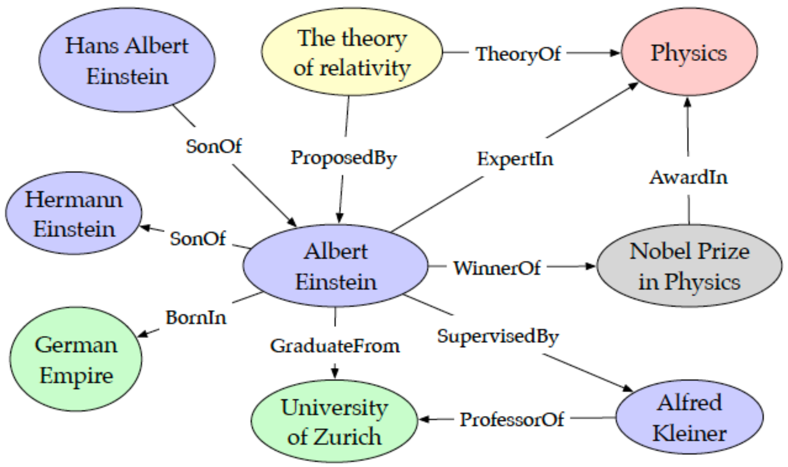
Outline

- Background and Motivation
- our Model: DisenKGAT
- Experiments
- Summary

Background and Motivation

- Knowledge Graphs (KGs) is a **structured representation** of facts, consisting of entities, relationships and semantic descriptions.
  - applied in dialogue generation, question answering, recommender systems.
  - **changed the paradigm** for numerous state-of-the-art natural language processing solutions.

(Albert Einstein, **BornIn**, German Empire)
(Albert Einstein, **SonOf**, Hermann Einstein)
(Albert Einstein, **GraduateFrom**, University of Zurich)
(Albert Einstein, **WinnerOf**, Nobel Prize in Physics)
(Albert Einstein, **ExpertIn**, Physics)
(Nobel Prize in Physics, **AwardIn**, Physics)
(The theory of relativity, **TheoryOf**, Physics)
(Albert Einstein, **SupervisedBy**, Alfred Kleiner)
(Alfred Kleiner, **ProfessorOf**, University of Zurich)
(The theory of relativity, **ProposedBy**, Albert Einstein)
(Hans Albert Einstein, **SonOf**, Albert Einstein)



An example of knowledge base and knowledge graph

Figure credit *to S. Ji's survey (A Survey on Knowledge Graphs: Representation, Acquisition and Applications)*

Background and Motivation

- Knowledge Graphs (KGs) is a **structured representation** of facts, consisting of entities, relationships and semantic descriptions.
  - Due to the constantly emerging new knowledge, they are still **far away from completeness**

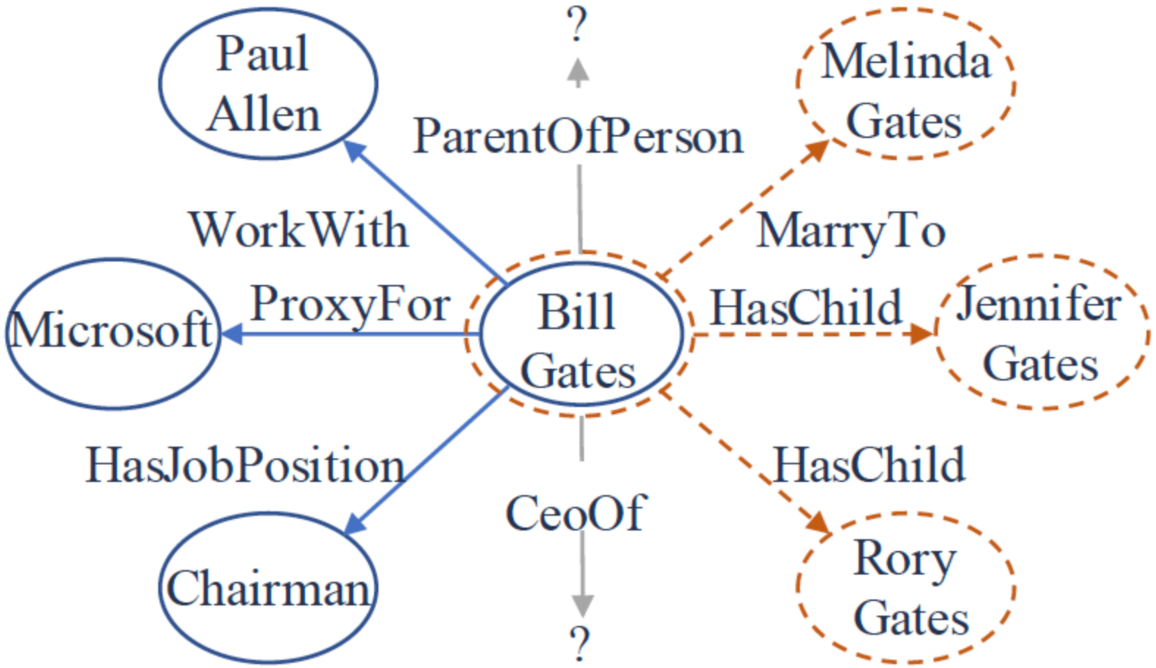


Figure credit *to J. sheng et al. 2020 Adaptive Attentional Network for Few-Shot Knowledge Graph Completion*

Knowledge Graph Embedding

- Distance-based model
  1. Inspired by the capability observed in Word2vec （$h + r \approx t$ ）.
  2. Gradually developed into various space (the complex space or Polar coordinates)

(transE [2013], transR [2016], RotatE [2019])

- Semantic models
  1. consider the KG as a 3D adjacency matrix
  2. score function is computed as a bilinear product $\emptyset(h, r, t) = h \times r \times t$

(DistMult [2015], ComplexE [2016])

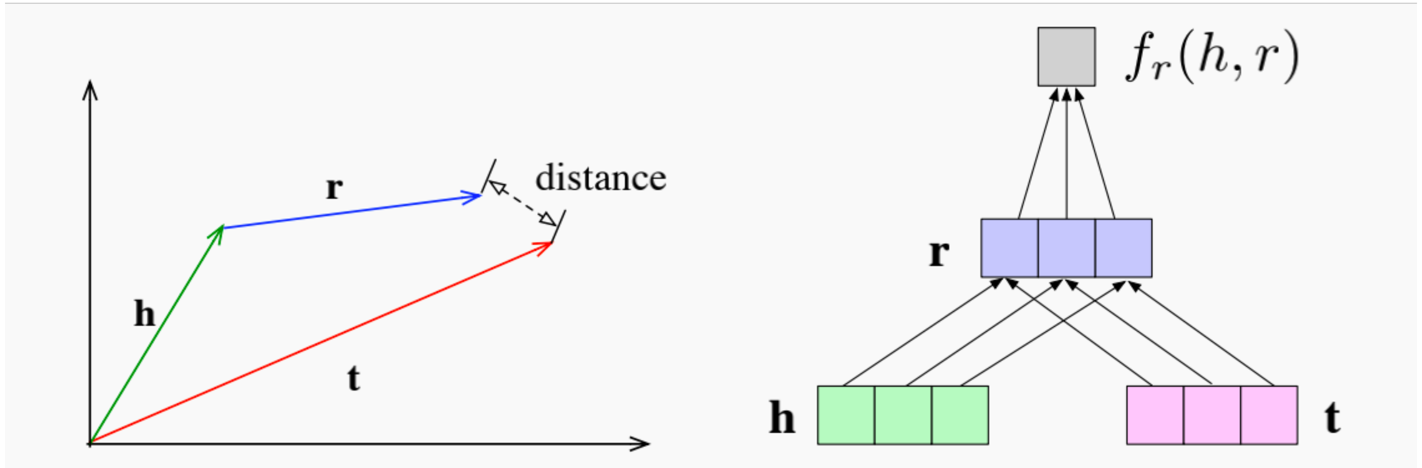
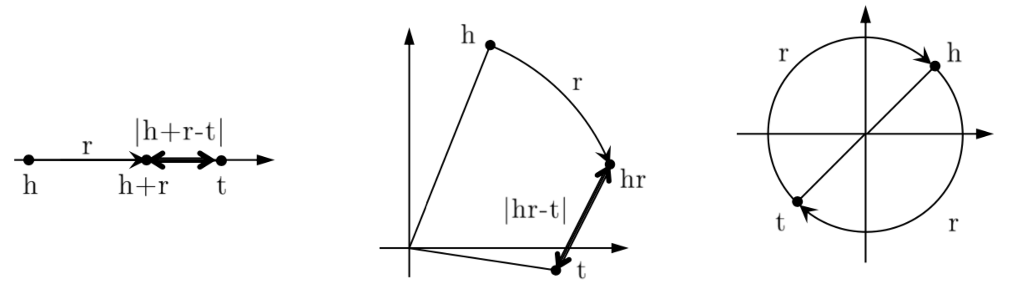


Figure credit *to S. Ji's survey (A Survey on Knowledge Graphs: Representation, Acquisition and Applications)*



(a) TransE models $\mathbf{r}$ as translation in real line.

(b) RotatE models $\mathbf{r}$ as rotation in complex plane.

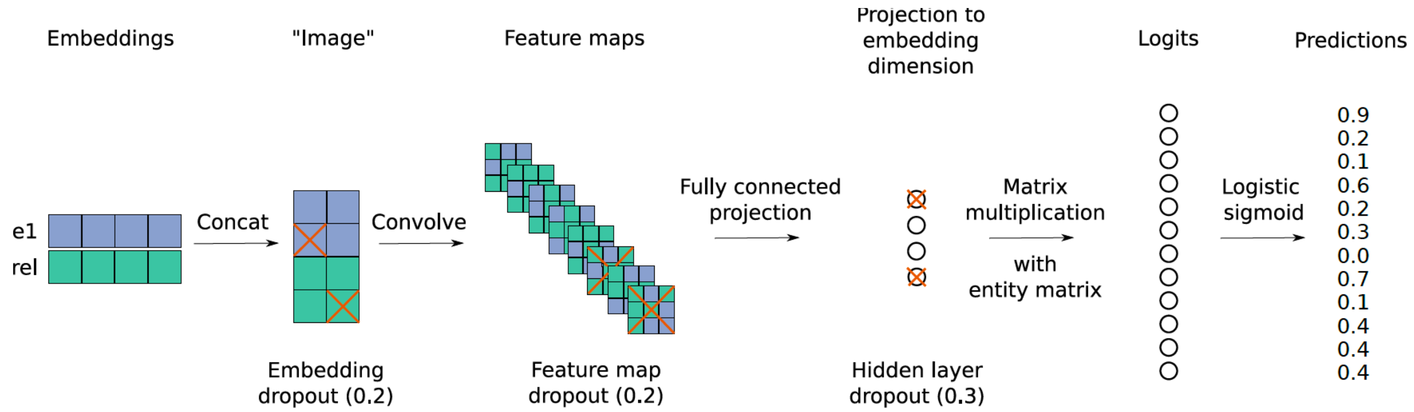
(c) RotatE: an example of modeling symmetric relations $\mathbf{r}$ with $r_i = -1$

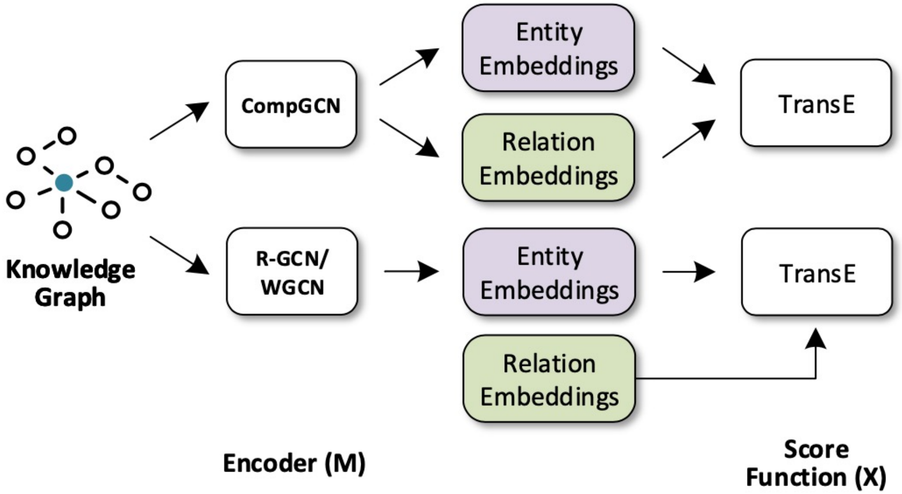
Figure credit *to Z. Sun et al. 2019 ICLR ROTATE: KNOWLEDGE GRAPH EMBEDDING BY RELATIONAL ROTATION IN COMPLEX SPACE*

Knowledge Graph Embedding

- Neural network based models
  1. leverage 2D convolution network to model the interaction.
  2. GCN constructs the encoder-decoder paradigm in knowledge graph completion.

(ConvE [2018], SACN [2019], CompGCN [2020])



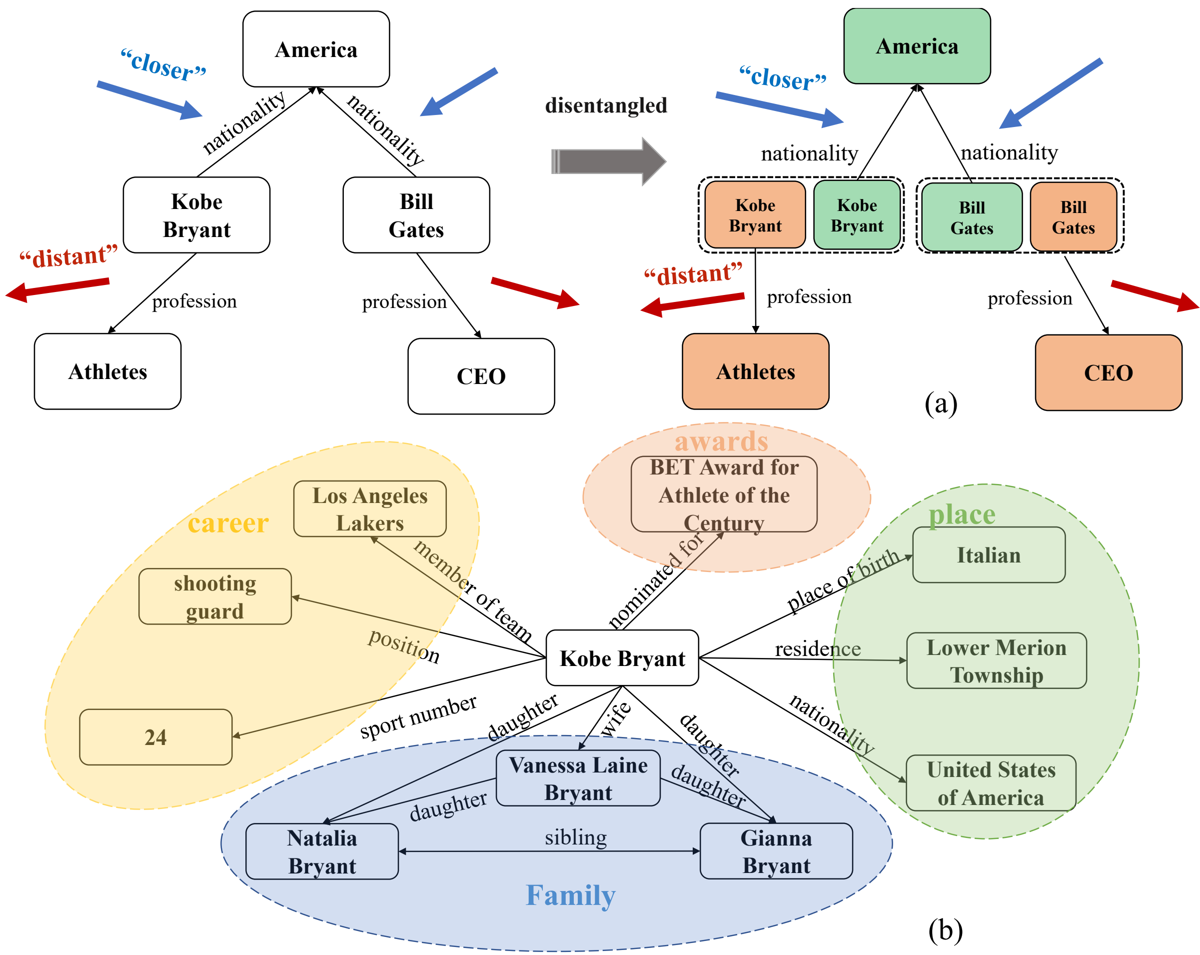


Define a score function and make sure

$$S_{pos} > S_{neg}$$

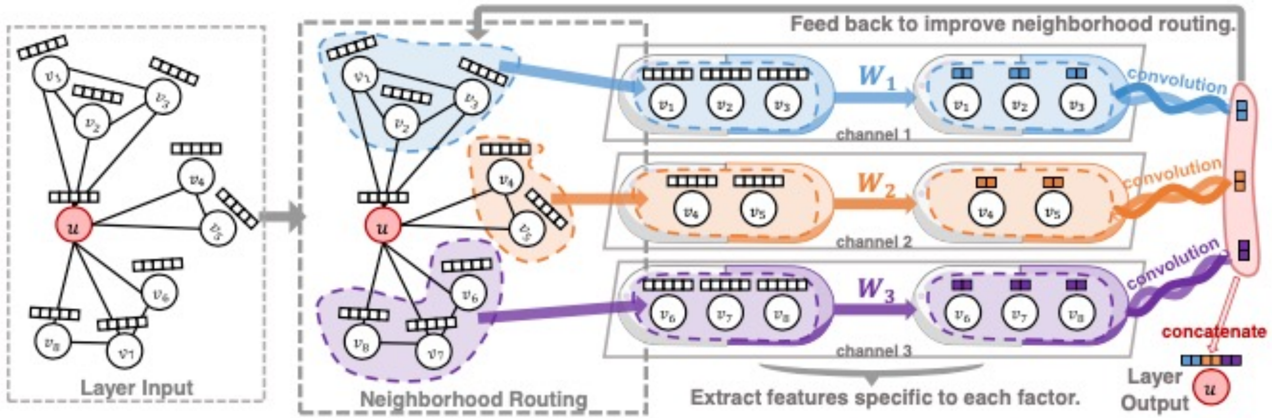
Drawbacks

- **Ignore the entanglement** of the latent factors behind the entity embeddings.
- The **static** representation fails to effectively model the critical relationship
- In the light of the above two points, these methods result in **low interpretability** and **non-robustness.**



Research Goal and Challenges

- **Goal**: provide a novel Disentangled Knowledge graph embedding framework which could predict adaptively according to the given scenario.
  - **Capture the latent factor[1].**
  - **Micro-disentanglement and Macro-disentanglement[2, 3].**



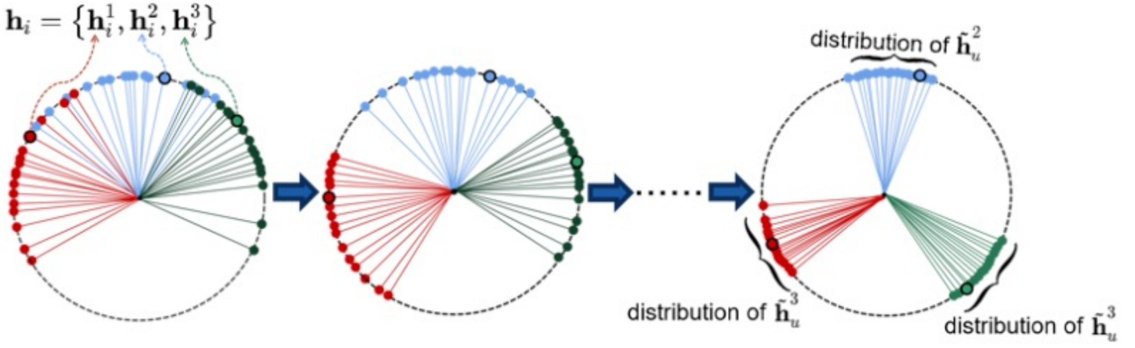
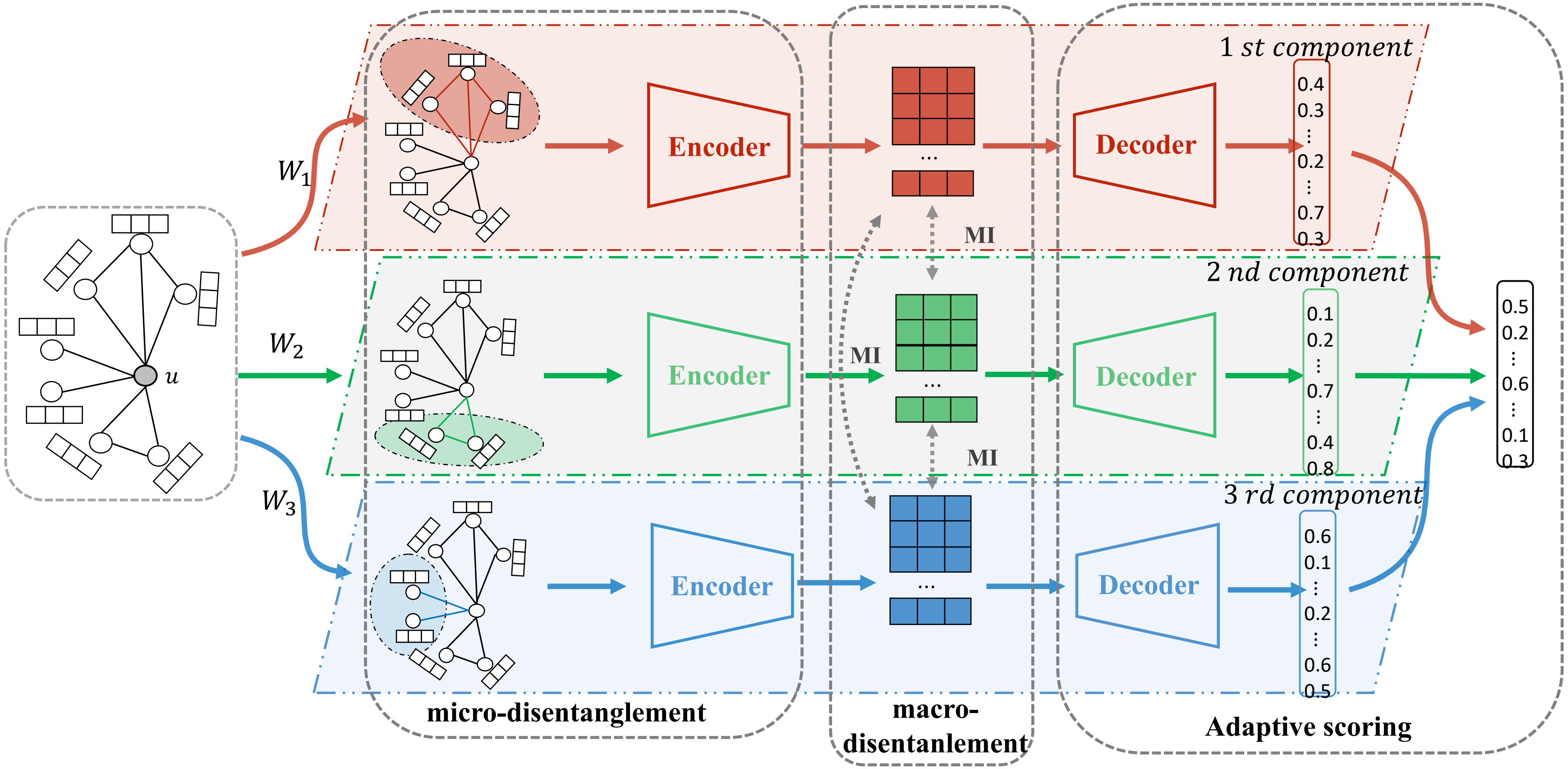


Figure credit *to J. Ma et al. Disentangled graph convolutional networks, ICML2019 and S. Zheng et al. Adversarial Graph Disentanglement*

[1]Jianxin Ma, et al. Disentangled graph convolutional networks. In ICML2019.
[2] Yanbei Liu, et al 2020. Independence promoted graph disentangled networks. In AAAI2020
[3] Shuai Zheng, et al. Adversarial Graph Disentanglement. 2021 arXiv preprint arXiv:2103.07295(2021)

Framework



An architecture overview of our DisenKGAT. The whole model contains three key modules: (1) relation-aware aggregation, (2) independence constraint, and (3) adaptive scoring.

Model

Disentangled Transformation

$$h_{u,k}^{0} = \sigma(W_k \cdot x_u) \longrightarrow$$ Normalization will hurt the final performance

Relation-aware Aggregation

$$m_{(v,k,r)} = \phi(h_{v,k}, h_r, \theta_r) \longrightarrow$$ **Explicitly combine crucial information-the edge relation**

$$\theta_r = W_r = diag(w_r)$$

$$\alpha_{(u,v,r)}^{k} = softmax\left(\left(e_{u,r}^{k}\right)^{T} \cdot e_{v,r}^{k}\right) = \frac{\exp\left(\left(e_{u,r}^{k}\right)^{T} \cdot e_{v,r}^{k}\right)}{\sum_{(v',r)\in\hat{N}(u)} \exp\left(\left(e_{u,r}^{k}\right)^{T} \cdot e_{v',r}^{k}\right)}$$

$$h_{u,k}^{l+1} = \sigma\left(\sum_{(v',r)\in\hat{N}(u)} \alpha_{(u,v,r)}^{k} \, \phi(h_{v,k}^{l}, h_r^{l}, \theta_r)\right)$$

the more similar the entity u and the neighbor v are in the k-th component in terms of their relation r , the more likely the factor k is to be the reason for the connection.

Independence Constraint

- Mutual information $\mathrm{I}(X,Z) = \mathbb{E}_{p(x,z)}\left[log\frac{p(x,z)}{p(x)p(z)}\right]$
- Mutual Information Minimization
  - **Nonlinear dependence** is crucial in complex Knowledge Graph.
  - utilize the contrastive log-ratio **upper bound** MI estimator[1]

$$\mathcal{L}_{mi} = \sum_i \sum_j \mathbb{E}_{(h_{u,i},h_{u,j})\sim p(h_{u,i},h_{u,j})}[logq(z_{u,i}|z_{u,j})] - \mathbb{E}_{(h_{u,i},h_{u',j})\sim p(h_{u,i})p(h_{u,j})}[logq(z_{u,i}|z_{u',j})]$$

  - leverage a variational distribution $q_\theta((h_{u,i}|h_{u,j})$ to approximate the real conditional one.

$$\mathcal{L}_{(h_{u,i},h_{u,j})} = \mathcal{D}_{kl}(p(h_{u,i}|h_{u,j})\,||q_\theta((h_{u,i}|h_{u,j}))$$

[1] Pengyu Chenge.et al. CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information. 2020ICML

Model

Adaptive Scoring

- Component-level prediction
  - ✓ compute the score for each candidate triplets (u, r, v)
  - ✓ take ConvE as an example

$$\boxed{\psi_{(u,r,v)}^{k}} = f\left(vec\left(f\left(\overline{h_{u,k}^{l}; h_{r}^{l}} * \omega\right)\right)W\right)h_{v,k}^{l}$$

- Relation-aware attentive fusion
  - ✓ the best-matched component representation should be **closer** to the given relation embedding.
  - ✓ $\boldsymbol{\theta_r}$ **is shared** with the relation-aware aggregation module.

Barack Obama	mother	Anne Dunham
Barack Obama	career	president
Barack Obama	Married_to	Michelle Obama
Michelle Obama	Daughter	Malia Ann Obama
Barack Obama	Daughter	???

$$\beta_{(u,r)}^{k} = softmax\left(\left(h_{u,k}^{L} \circ \theta_r\right)^T \cdot h_r^L\right)$$

$$= \frac{\exp\left(\left(h_{u,k}^{L} \circ \theta_r\right)^T \cdot h_r^L\right)}{\sum_{k'} \exp\left(\left(h_{u,k'}^{L} \circ \theta_r\right)^T \cdot h_r^L\right)}$$

$$\psi_{(u,r,v)}^{final} = \sum_{k} \beta_{(u,r)}^{k} \boxed{\psi_{(u,r,v)}^{k}}$$

Experiment

- ❑ How does DisenKGAT perform compared to existing aproaches, w.r.t. distance-based and semantic matching models?
- ❑ How do the critical components (e.g., relation-aware aggregation) contribute to DisenKGAT and how do different hyperparameters (e.g., factor number) affect DisenKGAT?
- ❑ Does DisenKGAT work robustly with other decoder modules?
- ❑ Can DisenKGAT give explanations of the benefits brought by the disentangled factors?

Baseline:

- ✓ Distance-based model (TransE, RotatE)
- ✓ Semantic models (DistMult, RESCAL)
- ✓ Neural network based models (ConvE, InteractE, SACN, ArcE, ReinceptionE, COMPGCN)

Data sets	$\|\mathcal{E}\|$	$\|\mathcal{R}\|$	Triplets		
			Train	Valid	Test
FB15k-237	14,541	237	272,114	17,535	20,466
WN18RR	40,943	11	86,835	3,034	3,134

Performance

Model	FB15k-237					WN18RR				
	MRR	MR	Hits@1	Hit@3	Hit@10	MRR	MR	Hits@1	Hit@3	Hit@10
TransE [5]	0.294	357	-	-	0.465	0.226	3384	-	-	0.501
Distmult [42]	0.241	254	0.155	0.263	0.419	0.43	5110	0.39	0.44	0.49
ConvE [8]	0.325	244	0.237	0.356	0.501	0.43	4187	0.40	0.44	0.52
RotatE [31]	0.338	177	0.241	0.375	0.533	0.476	3340	0.428	0.492	0.571
SACN [29]	0.35	-	0.261	0.39	0.54	0.47	-	0.43	0.48	0.54
InteractE [34]	0.354	**172**	0.263	-	0.535	0.463	5202	0.43	-	0.528
MuRE [2]	0.336	-	0.245	0.370	0.521	0.465	-	0.436	0.487	0.554
COMPGCN [35]	0.355	197	0.264	0.39	0.535	0.479	3533	**0.443**	0.494	0.546
AcrE [27]	0.358	-	0.266	0.393	0.545	0.459	-	0.422	0.473	0.532
ReInceptionE [41]	0.349	173	-	-	0.528	0.483	1894	-	-	0.582
DisenKGAT	**0.368**	179	**0.275**	**0.407**	**0.553**	**0.486**	**1504**	0.441	**0.502**	**0.578**

- DisenKGAT achieve a considerable improvement on FB15k-237 which includes 237 relations
- WN18RR only contain 11 relation types.

Performance

		RotatE		WGCN		COMPGCN		DisenKGAT	
		MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10
Head Pred	1-1	0.498	0.593	0.422	0.547	0.457	0.604	**0.501**	**0.625**
	1-N	0.092	0.174	0.093	0.187	0.112	0.190	**0.128**	**0.248**
	N-1	0.471	0.674	0.454	0.647	0.471	0.656	**0.486**	**0.659**
	N-N	0.261	0.476	0.261	0.459	0.275	0.474	**0.291**	**0.496**
Tail Pred	1-1	0.484	0.578	0.406	0.531	0.453	0.589	**0.499**	**0.641**
	1-N	0.749	0.674	0.771	0.875	0.779	0.885	**0.789**	**0.889**
	N-1	0.074	0.138	0.068	0.139	0.076	0.151	**0.086**	**0.180**
	N-N	0.364	0.608	0.385	0.607	0.395	0.616	**0.402**	**0.629**

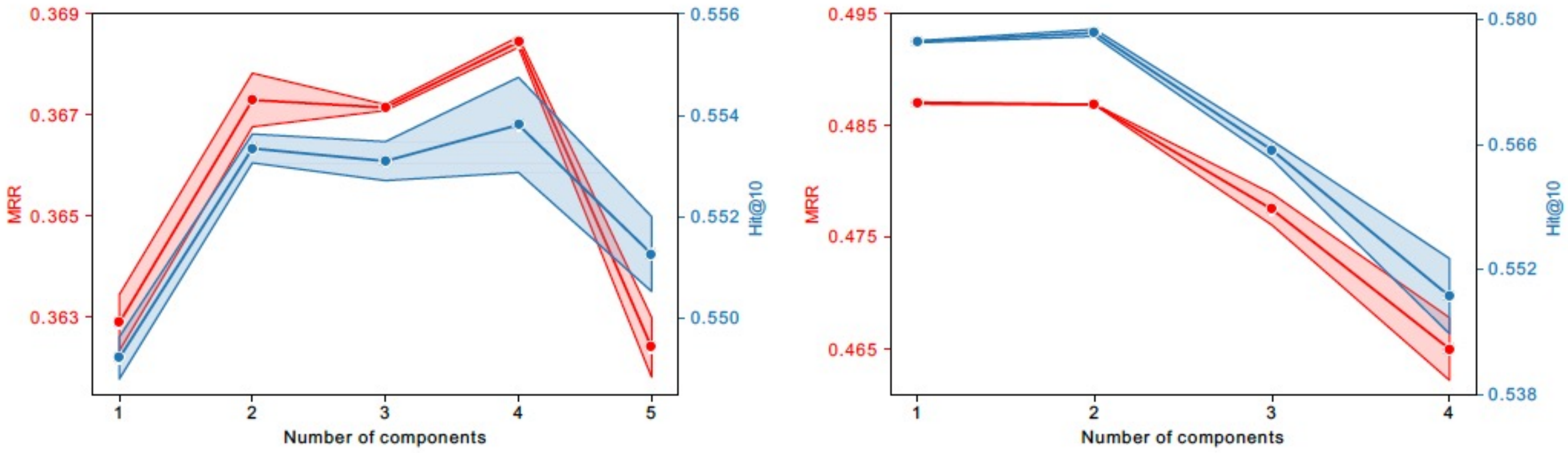
- GNN-based models (W-GCN, COMPGCN) are superior to RotatE on **complex relation types** (1-N, N-1, N-N)
- RotatE outperforms W-GCN and COMPGCN on **simple relation** (1-1) including symmetry/antisymmetry, composition, and inversion.
- Our model outperforms other models by a large margin in **both simple and complex relations**.

Performance

model	MRR	MR	Hits@1	Hits@3	Hits@10
w/o micro	0.355	197	0.265	0.392	0.534
w/o macro	0.356	303	0.263	0.392	0.542
w/o HSIC	0.352	259	0.263	0.387	0.527
DisenKGAT	0.368	179	0.275	0.407	0.553

- Without micro-disentanglement, the model **degrades** to vanilla GCN-based model.
- Without macro-disentanglement, each component is prone to **entangle again**!
- HSIC is not suitable for more complex heterogeneous graphs.

Performance



(a) FB1515k-237 (b) WN18RR

- K=1, it degrades into a normal GNN-based model but coupling with attention aggregation and relation-aware mapping.
- K>4, it makes some topics too fine-grained to carry **crucial information**.
- the performance on WN18RR collapses significantly in term to its simple meaning.

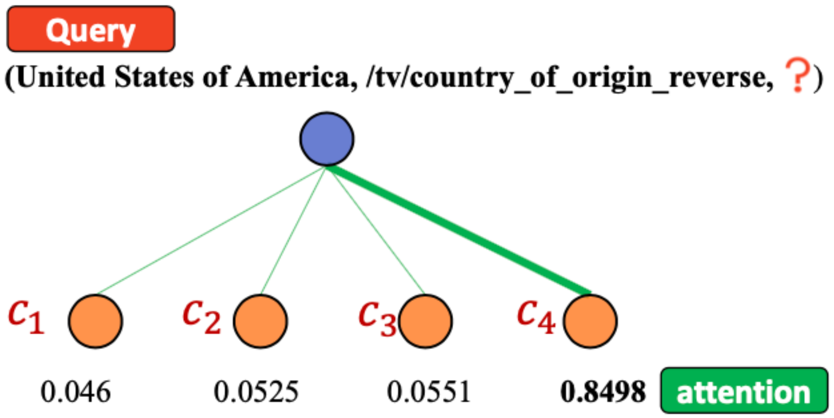
Performance

Scoring Function(=X)→	TransE			DistMult			ConvE		
Methods↓	MRR	MR	H@10	MRR	MR	H@10	MRR	MR	H@10
X	0.294	357	0.465	0.241	354	0.419	0.325	244	0.501
X+D-GCN	0.299	351	0.469	0.321	225	0.497	0.344	200	0.524
X+W-GCN	0.264	1520	0.444	0.324	229	0.504	0.244	201	0.525
X+COMPGCN(sub)	0.335	194	0.514	0.336	231	0.513	0.352	199	0.530
X+COMPGCN(Mult)	0.337	233	0.515	0.338	200	0.518	0.353	216	0.532
X+COMPGCN(Corr)	0.336	214	0.518	0.335	227	0.514	0.355	197	0.535
X+DisenKGAT(sub)	0.334	<u>183</u>	0.51	0.346	<u>196</u>	0.531	0.358	181	0.543
X+DisenKGAT(Mult)	<u>0.342</u>	**170**	<u>0.524</u>	<u>0.353</u>	**184**	<u>0.536</u>	<u>0.364</u>	**171**	<u>0.550</u>
X+DisenKGAT(Corr)	0.338	203	0.520	0.341	200	0.528	0.359	189	0.541
X+DisenKGAT(Cross)	**0.343**	187	**0.526**	**0.354**	204	**0.540**	**0.368**	<u>179</u>	**0.553**

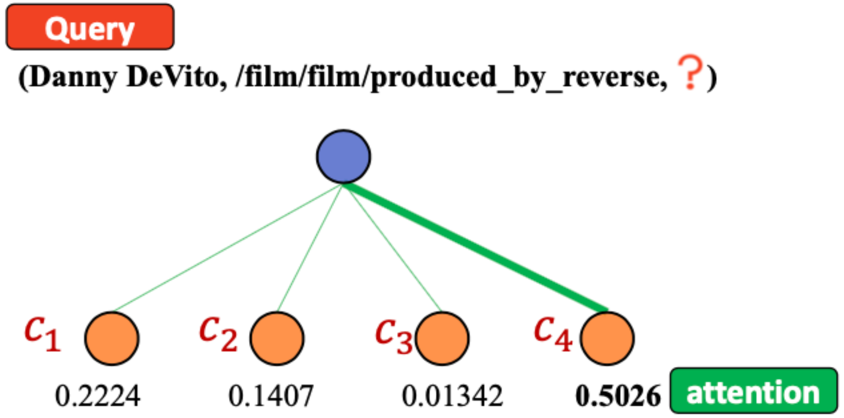
- Subtraction (Sub): $\emptyset(e_s, e_r, \theta) = (\theta_r \cdot e_s) - e_r$
- Multiplication(Mult): $\emptyset(e_s, e_r, \theta) = (\theta_r \cdot e_s) \circ e_r$
- Circular correlation(Corr): $\emptyset(e_s, e_r, \theta) = (\theta_r \cdot e_s) \star e_r$
- Crossover Interaction(Cross): $\phi(e_s, e_r, \theta) = \theta_r \cdot e_s + \theta_r(e_s \circ e_r)$

Performance

	Top 2 neighbors in each component	Score
c_1	(Olympics, 1924 Summer Olympics)	0.136
	(Olympics, 1900 Summer Olympics)	0.128
c_2	(location/contains, Washington metropolitan area)	0.0018
	(location/contains, LaGuardia Airport)	0.0016
c_3	(location/contains, University of Maryland).	0.0027
	(location/contains, Southern Methodist University)	0.0026
c_4	(tv/country_of_origin_reverse, All My children)	0.0033
	(tv/country_of_origin_reverse, Backstairs at the White House	0.0024



	Top 2 neighbors in each component	Score
c_1	(person/profession, voice actor)	0.091
	(person/profession, television director)	0.090
c_2	(award_nominee, gender, Kim Basinger)	0.012
	(award_nominee, Michael Shamberg)	0.011
c_3	(person/gender, male organism)	0.024
	(person/type_of_union, marriage)	0.020
c_4	(film/produced_by_reverse, Get Shortly)	0.039
	(film/produced_by_reverse, Be Cool)	0.032



- Construct a **distinguishable clusters** potentially.
- Topic or cluster in each component of various entities should be **shared all the time**.

Summary

- We propose a novel Disentangled Knowledge attention network, DisenKGAT.

- We take the micro-macro disentanglement into consideration simultaneously.

- We look forward to explore a more general disentangled framework that could adapt to more complex scenarios.

Future

- More flexible disentanglement combine with **adaptive K**.
- Disentanglement in more research areas.
- Contrastive learning in KG.