Graph-based Machine Learning

Analogy for Knowledge Base Completion

(H Liu et al. ICML 2017)

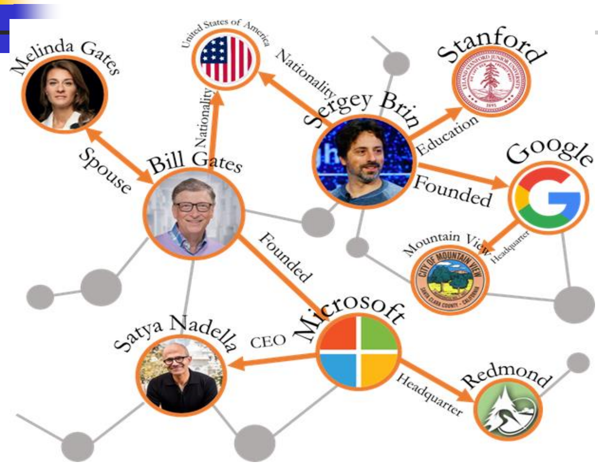
1

Outline

- Recap of KG completion methods (previous lecture)
- Analogy Modeling (Hanxiao Liu, et al., ICML 2017)

2

Knowledge Graph (KG)



- **KG completion task**
  - (Sergey, Spouse, ?)
  - (Nadella, ?, Gates)
- **Down stream tasks**
  - Knowledge-enhanced QA, IR, Classification, KG-alignment, etc.
- **Research**
  - Neural embedding

3

Expressiveness of Different Methods
(Zhiqing Sun et al., ICLR 2019)

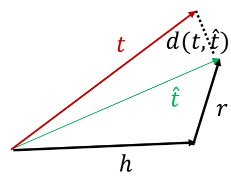
Model	Score Function	Symmetry	Antisymmetry	Inversion	Composition	Analogy
SE	$-\|\boldsymbol{W}_{r,1}\mathbf{h} - \boldsymbol{W}_{r,2}\mathbf{t}\|$	✗	✗	✗	✗	
TransE	$-\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|$	✗	✓	✓	✓	
TransX	$-\|g_{r,1}(\mathbf{h}) + \mathbf{r} - g_{r,2}(\mathbf{t})\|$	✓	✓	✗	✗	
DistMult	$\langle \mathbf{h}, \mathbf{r}, \mathbf{t} \rangle$	✓	✗	✗	✗	
ComplEx	$\mathrm{Re}(\langle \mathbf{h}, \mathbf{r}, \mathbf{t} \rangle)$	✓	✓	✓	✗	
RotatE	$-\|\mathbf{h} \circ \mathbf{r} - \mathbf{t}\|$	✓	✓	✓	✓	
Analogy (ICML'2017)	$-log\sigma(y\boldsymbol{h}^T W_r \boldsymbol{t})$	✓	✓	✓	?	✓

- What method is missing above?
- How the above methods related to KE-GCN?

4

Recap of TransE (Bordels et al., NIPS 2013)

- Learning a real vector embedding for each entity and relation
- Predicting the missing element in $(h, r, ?)$ by calculating $f(\boldsymbol{h}, \boldsymbol{r}) = \boldsymbol{h} + \boldsymbol{r} \triangleq \hat{\boldsymbol{t}}$
- Minimizing distance $d(\boldsymbol{t}, \hat{\boldsymbol{t}}) = \|\boldsymbol{t} - \hat{\boldsymbol{t}}\| = \|\boldsymbol{h} + \boldsymbol{r} - \boldsymbol{t}\|$ during training (iterative optimization of embedding vectors)

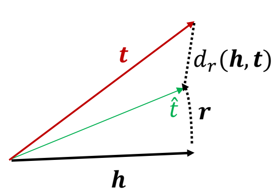


Vector $\boldsymbol{r}$ is added to vector $\boldsymbol{h}$.

5

Recap of RotatE (Zhiqing Sun et al., ICLR 2019)

- Learning a complex vector embedding for each entity and relation
- Predicting the missing element in $(h, r, ?)$ by calculating $f(\boldsymbol{h}, \boldsymbol{r}) = \boldsymbol{h} \circ \boldsymbol{r} \triangleq \hat{\boldsymbol{t}}$
- Minimizing distance $d_r(\boldsymbol{h}, \boldsymbol{t}) = \|\boldsymbol{h} \circ \boldsymbol{r} - \boldsymbol{t}\|$ during training



$\boldsymbol{r}$ is a unit-length rotation operator

6

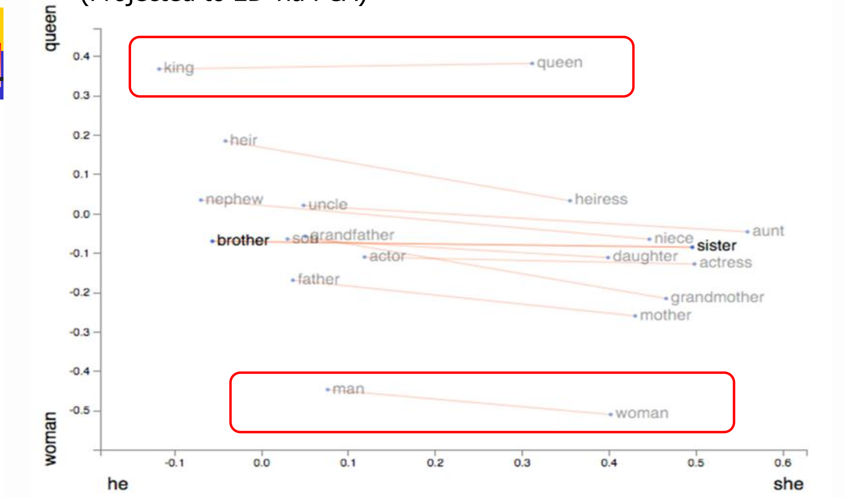
Important Types of KG Relations

- Symmetric (h and t) (H Liu, ICML 2017)
  - $\phi(h, r, t) = \phi(t, r, h)$, e.g., marriage, divorce
- Antisymmetric (h and t) (H Liu, ICML 2017)
  - $\phi(h, r, t) \gg \phi(t, r, h)$, e.g., a-parent-of
- Inversive (r and r') (H Liu, ICML 2017)
  - $\phi(h, r, t) = \phi(h, r', t)$, e.g., hypernym (r) vs. hyponym (r')
- Compositional (or "transitive") (r and r') (Z Sun, ICLR 2019)
  - $\phi(a, r, b) \times \phi(b, r', c) = \phi(a, r \circ r', c)$, e.g., my mother's husband is my father
- **Commutative** (r and r'): $r \circ r' = r' \circ r$ (H Liu, ICML 2017)
  - $\phi(a, r \circ r', d) = \phi(a, r' \circ r, d)$, e.g., king to queen as man to women

Outline

- Recap of KG completion methods (previous lecture)
- Analogical inference for multi-relational embeddings (H Liu, et al., ICML 2017)
  - Mathematical modeling of analogy with differentiable optimization
  - A unified framework subsuming several representative methods
  - Fast algorithm for linear scalability

Observed Analogy in Masculine-Feminine Vectors by GloVe
(Projected to 2D via PCA)



http://p.migdal.pl/2017/01/06/king-man-woman-queen-why.html

9

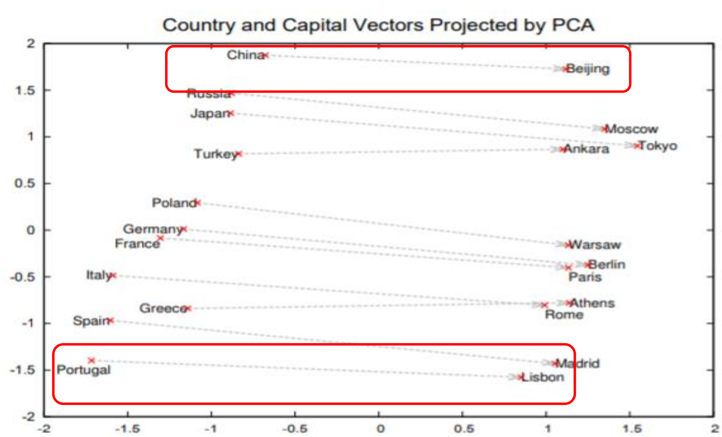


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

10

Geometric Property of Analogy

- If two systems form an analogy, then understanding one of them would help the understanding of the other.

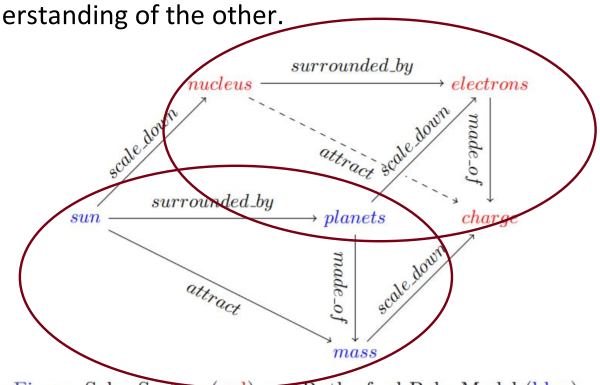


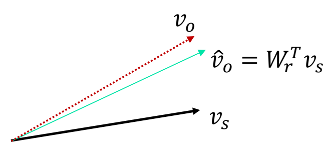
Figure: Solar System (red) v.s. Rutherford-Bohr Model (blue).

Notation

- Triplet (s, r, o) is any subject-relation-object (or head-relation-tail) combinations;
- Labeled training set $\mathcal{D} = \{((s, r, o), y)\}$ with $y = \pm 1$ as the labels (positive vs. negative);
- Vectors $v_s \in R^d$ and $v_o \in R^d$ are the learned embeddings for s and o, respectively;
- Matric $W_r \in R^{d \times d}$ is the learned embedding of relation r;
- Boldfaced $\boldsymbol{v}$ is the collection of the vector embeddings for all entities in $\mathcal{D}$;
- Boldfaced ***W*** is the collection of the matrix embeddings for all relations in $\mathcal{D}$.

Treating Relation as a Linear Translation Operator

- Linear transformation (via W_r^T) from v_s to v_o



- For each semantically valid triplet (s, r, o), we want

$$W_r^T v_s = \hat{v}_o \approx v_o \tag{1}$$

- Scoring Function (higher is better):

$$\phi(s, r, o) = \langle W_r^T v_s, v_o \rangle = v_s^T W_r v_o \tag{2}$$

Constrained Optimization

$$\min_{\boldsymbol{v}, \boldsymbol{W}} \mathbb{E}_{((s,r,o),y)\in\mathcal{D}} l\big(\phi_{\boldsymbol{v},\boldsymbol{W}}(s, r, o), y\big) \tag{3}$$

$$s.t. \quad W_r W_r^T = W_r^T W_r, \ \forall r \tag{4}$$

$$W_r W_{r'} = W_{r'} W_r, \ \forall r, r' \tag{5}$$

- Formula (4) restricts the matrices to be in the family of normal matrices;
- Formula (5) restricts the matrices to have the commutative property;
- Together they define the desirable properties of analogy.

Normal Matrices (Formula 4)

- Def. A linear transformation W_r which satisfies

$$W_r^T W_r = W_r\, W_r^T \qquad (4)$$

- Properties ("well-behaved")
  - Symmetric: $W_r = W_r^T$ (4a)
  - Anti-symmetric: $W_r = -W_r^T$ (4b)
  - Bijective (Orthogonal): $W_r^T W_r = W_r W_r^T = \mathrm{I}$, or $W_r^T = W_r^{-1}$ (4c)
  - ...
- The above cases can be easily verified (try by yourself)!

Symmetric Relationship with $W_r = W_r^T$ (4a)

- Example: "Alex is married to Bob" implies "Bob is married to Alex", and vise versa.
- We need to show $\phi(s,r,o) = \phi(o,r,s)$ if $W_r = W_r^T$.
- Proof

$\phi(s,r,o) = \langle W_r^T v_s, v_o\rangle = v_s^T W_r v_o$ (by the scoring function def.)

$\phi(o,r,s) = \langle W_r^T v_o, v_s\rangle = \langle v_s, W_r^T v_o\rangle$ (dot-product is symmetric)

$= v_s^T W_r^T v_o = v_s^T W_r v_o$ (substituting W_r^T by W_r)

$\therefore \quad \phi(s,r,o) = \phi(s,r,o) \quad ■$

Anti-symmetric Relationship with $W_r = -W_r^T$ (4b)

- Example: "Alex is the mother of Bob" implies "Bob is not the mother of Alex".
- We need to show $\phi(s,r,o) = -\,\phi(o,r,s)$ if $W_r \; = -W_r^T$.
- Proof

$\phi(s,r,o) = \langle W_r^T v_s, v_o \rangle \; = v_s^T W_r v_o$ (by the scoring function def.)

$\phi(o,r,s) = \langle W_r^T v_o, v_s \rangle = \langle v_s, W_r^T v_o \rangle$ (dot-product is symmetric)

$= v_s^T W_r^T v_o = -\, v_s^T W_r v_o$ (substituting W_r^T by $-W_r$)

$\therefore \quad \phi(s,r,o) = -\phi(s,r,o) \quad \blacksquare$

17

Bijective Relationship with $W_r^T \; = \; W_r^{-1}$ (4c)

- Bijective means 1-to-1 mapping, i.e., a stronger condition than symmetric.
- Example: "My (unique) cmu-andrew ID is yiming".
- Consider a rotation matrix W_θ and the inverse rotation

$$W_\theta = \begin{bmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix} \qquad W_{-\theta} = \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix}$$

- Clearly, we have $W_\theta^{\mathrm{T}} = W_{-\theta}$ and

$$W_\theta \, W_{-\theta} = \begin{bmatrix} cos\theta cos\theta + sin\theta sin\theta & -cos\theta sin\theta + sin\theta cos\theta \\ -sin\theta cos\theta + cos\theta sin\theta & sin\theta sin\theta + cos\theta cos\theta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

- That is

$W_{-\theta} = W_\theta^{-1} = W_\theta^{\mathrm{T}}$ (the inverse of W_θ always exists, which is W_θ^{T}.)

18

Bijective in Analogy vs. "Inversive" in RotatE

- Relation θ being bijective (H Liu, ICLM 2017) means $W_\theta W_\theta^T = I$.
  - The inverse of W_θ always exists, which is $W_\theta^T = W_{-\theta}$.
- Relations r_1 and r_2 being inversive (Z Sun, ICLR 2019) means $\boldsymbol{v}_{r1} \circ \boldsymbol{v}_{r2} = \mathbf{1}$.
  - The corresponding element-wise rotations θ_1 and θ_2 always satisfy $\theta_1 = -\theta_2$.
- We are just talking about the same pattern of relations in different words, i.e., the rotation by θ and the rotation inverse by $-\theta$.

19

Commutativity (Formula 5)

- Commutativity is a necessary condition for analogy. Mathematically, we have

$$W_r W_{r'} = W_{r'} W_r, \quad \forall r, r' \qquad (5)$$

- Equivalently, we can express this property as

$$r \circ r' = r' \circ r \qquad \forall r, r'$$

- Geometrically, r, r' define a parallelogram

$$\begin{tikzcd} a \arrow[r, "r"] \arrow[d, "r'"'] & b \arrow[d, "r'"] \\ c \arrow[r, "r"] & d \end{tikzcd}$$

20

Concrete Example of Analogical Structure

- **Well-known example**
  - "man is to king as woman is to queen"
- **Abstract notion**
  - "a is to b as c is to d"

$$\begin{tikzcd} a \arrow[r, "r"] \arrow[dr, "r'"] & b \arrow[dr, "r'"] & \\ & c \arrow[r, "r"] & d \end{tikzcd}$$

- **Geometrically**, consider that $r$and r' define a parallelogram, with

$$\phi(a, r \circ r', d) = \phi(a, r' \circ r, d)$$

In words, analogy is defined by the commutativity of relations.

Computational Challenge

$$\min_{\boldsymbol{v},\boldsymbol{W}} \mathbb{E}_{((s,r,o),y)\in\mathcal{D}}\, l\big(\phi_{\boldsymbol{v},\boldsymbol{W}}(s,r,o),y\big)$$

$$s.t. \quad W_r W_r^T = W_r^T W_r \;, \forall r \qquad (4)$$

$$W_r W_{r'} = W_{r'} W_r, \;\; \forall r, r' \qquad (5)$$

- **Good news**: Our objective is differentiable (supporting gradient descend).
- **Bad news**: The number of constraints can be very large (quadratic in the # of relations), and each relation has a dense-matrix representation (for $v \in \mathbb{R}^d$, we have $W_r \in \mathbb{R}^{d\times d}$).

The Remedy?

Lemma 4.1. *(Wilkinson & Wilkinson, 1965) For any real normal matrix A, there exists a real orthogonal matrix Q and a block-diagonal matrix B such that $A = QBQ^\top$, where each diagonal block of B is either (1) A real scalar, or (2) A 2-dimensional real matrix in the form of* $\begin{bmatrix} x & -y \\ y & x \end{bmatrix}$, *where both x, y are real scalars.*

The lemma suggests any real normal matrix can be block-diagonalized into an almost-diagonal canonical form.

The Magic?

- We can use $W_r = Q^T B_r Q$ to rewrite our scoring functions as

$$\forall r,\ \phi(s,r,o) = v_s^T W_r v_o = \underbrace{v_s^T Q^T}_{v_s'^T} B_r \underbrace{Q v_0}_{v_0'} = v_s'^T B_r v_0'$$

where B_r has 1×1 or 2×2 non-zero diagonal blocks and zero's anywhere else.

- Now, we can solve the optimization problem for $\{v'\}$ and $\{B_r\}$ instead of $\{v\}$ and $\{W_r\}$ in the original objective function.

Replacing dense w_r by sparse w_r'

- Original Objective

$$\min_{\boldsymbol{v},\boldsymbol{W}} \mathbb{E}_{((s,r,o),y)\in\mathcal{D}}\, l\big(\phi_{\boldsymbol{v},\boldsymbol{W}}(s,r,o),y\big)$$

- New Equivalent Objective

$$\min_{\boldsymbol{v}',\boldsymbol{B}} \mathbb{E}_{((s,r,o),y)\in\mathcal{D}}\, l\big(\phi_{\boldsymbol{v}',\boldsymbol{B}_r}(s,r,o),y\big)$$

where each $B_r \in \boldsymbol{B}$ is a block-diagonal whose block sizes are bounded by 2.

- Time/Space Saving

$O(d^2) \rightarrow O(d)$ where d is the embedding size of vectors (for entities).

Implementation Details

- Use logistic loss

$$l(\phi(s,r,o),y) = -log\sigma\big(y\phi(s,r,o)\big) \quad y \in \{\pm 1\}$$

- Optimization algorithm
  - Asynchronous AdaGrad
- Negative training instances
  - For each valid (s,r,o), generate the negative examples (s',r,o), (s,r',o) and (s,r,o') by corrupting s, r, o, respectively.

Analogy subsumes other well-known methods

- Multiplicative Embedding (**DistMult** by Yang et al. CoRR 2014)

$$\phi(s,r,o) = \langle v_s, v_r, v_o \rangle \quad \text{where } v_s, v_r, v_o \in \mathbb{R}^d$$

  - equivalent to Analogy by setting $W_r \stackrel{\text{def}}{=} diag(v_r)$ as a special case

- Complex Embedding (**ComplEx** by Trouillon et al, ICML 2016)

$$\phi(s,r,o) = RealPart\langle v_s, v_r, \overline{v_o}\rangle \quad \text{where } v_s, v_r, v_o \in \mathbb{C}^d$$

  where <., ., .> denote the generalized dot-product.

  - The solution can be fully recovered by Analogy with embedding size of $2d$ because any complex number $a + bj$ is isomorphic to the 2×2 matrix $\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$.

Analogy subsumes other well-known methods (cont'd)

- Holographic Embeddings (**HolE** by Nickel et al., AAAI 2016)

$$\phi(s,r,o) = \langle v_s, v_r * v_o\rangle$$

  where $v_s, v_r, v_o \in \mathbb{R}^d$ and * denotes circular correlation.

- This is equivalent to solving

$$\phi(s,r,o) = RealPart(\langle v_s, v_r, \overline{v_o}\rangle)$$

  where $v_s, v_r, v_o \in FFT(\mathbb{R}^d) \in \mathbb{C}^d$ are the Fast Fourier Transform of real vectors.

- Hence, HolE is a restricted case of ComplEx and ANALOGY.

Evaluation Results (Hanxiao Liu et al., ICML 2017)

- Use Mean Reciprocal Rank (MRR) and Hits@k as the metrics
- Benchmark datasets of FreeBase-15K and WordNet-18

	WN18				FB15K			
Models	MRR (filt.)	MRR (raw)	Hits@1 (filt.)	Hits@3 (filt.)	MRR (filt.)	MRR (raw)	Hits@1 (filt.)	Hits@3 (filt.)
RESCAL	89.0	60.3	84.2	90.4	35.4	18.9	23.5	40.9
TransE	45.4	33.5	8.9	82.3	38.0	22.1	23.1	47.2
DistMult	82.2	53.2	72.8	91.4	65.4	24.2	54.6	73.3
HolE	93.8	61.6	93.0	94.5	52.4	23.2	40.2	61.3
ComplEx	94.1	58.7	93.6	94.5	69.2	24.2	59.9	75.9
Our ANALOGY	94.2	65.7	93.9	94.4	72.5	25.3	64.6	78.5

Results of ICML 2017 (H Liu) vs. ICLR 2019 (Z Sun)

- Results in Mean Reciprocal Rank (MRR)

Models	WN18 (2017)	WN18 (2019)	FB15k (2017)	FB15k (2019)
RESCAL	89	-	-	-
TransE	45.4	49.5	38.0	46.3
DistMult	82.2	79.7	65.4	79.8
HolE	93.8	93.8	52.4	52.4
ComplEx	94.1	94.1	69.2	69.2
ANALOGY	**94.2**	-	**72.5**	
ConvE	-	94.3		65.7
RotatE	-	**94.9**		**79.9**

Concluding Remarks

- Analogy can be formulated geometrically in a real vector space, supporting differentiable optimization for KG embedding with linear scalability.
- It provides a unified framework for several representative KG embedding methods.
- **Limitation:** Cannot model compositional relations? (Beyond the scope of this lecture)
- **Connection/difference from KE-GCN**
  - All the KG completion methods are designed without consideration of downstream tasks, but KE-GCN is.

KG-embedding Methods vs. KE-GCN

- KG-embedding methods (e.g., Analogy)

$$\min_{\boldsymbol{v},\boldsymbol{W}} \mathbb{E}_{((s,r,o),y)\in\mathcal{D}} l\big(\phi_{\boldsymbol{v},\boldsymbol{W}}(s,r,o),y\big)$$

- KE-GCN for multi-class classification

$$\mathcal{L} = -\sum_{(X_i,Y_i)\in\mathcal{D}_l} \sum_{j=1}^{K} Y_{ij} \ln \hat{Y}_{ij}$$

- KE-GCN for cross-language KG Alignment

$$\mathcal{L} = -\sum_{(u,v)\in S} \sum_{(u',v')\in S'} \left[\|h_u - h_v\|_1 - \|h_{u'} - h_{v'}\|_1 + \gamma \right]_+$$

Is the winning method for KG-completion necessarily be the best choice for down-stream tasks?

Table 11: Entity classification accuracy results over 5 different runs on AM and WN datasets by incorporating different knowledge graph embedding methods into our model.

KE-GCN (X)	AM	WN
X = TransE	**91.2 ± 0.2**	**57.8 ± 0.5**
X = TransH	90.5 ± 0.3	57.4 ± 0.3
X = DistMult	89.5 ± 0.4	56.4 ± 0.1
X = TransD	90.1 ± 0.2	57.1 ± 0.2
X = RotatE	90.6 ± 0.4	56.6 ± 0.3
X = QuatE	91.0 ± 0.4	56.9 ± 0.3

Table 4: Knowledge graph entity alignment results over 5 different runs on $DBP_{ZH\text{-}EN}$ by incorporating different knowledge graph embedding methods into our model.

KE-GCN (X)	MRR	H@1	H@10
X = TransE	0.648 ± 0.003	54.3 ± 0.3	83.4 ± 0.3
X = TransH	0.650 ± 0.003	54.3 ± 0.4	**84.4 ± 0.3**
X = DistMult	0.621 ± 0.003	52.0 ± 0.4	80.3 ± 0.4
X = TransD	0.635 ± 0.003	53.1 ± 0.3	82.7 ± 0.4
X = RotatE	0.653 ± 0.004	54.9 ± 0.4	83.8 ± 0.4
X = QuatE	**0.664 ± 0.004**	**56.2 ± 0.4**	84.2 ± 0.4

33

References

- Hanxiao Liu, Yuexin Wu, Yiming Yang. Analogical inference for multi-relational embeddings. ICML 2017.
- Zhiqing Sun, Zhi-hone Deng, Jian-yun Nie, Jian Tang. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. ICLR 2019.

34