LinkNBed: Multi-Graph Representation Learning with Entity Linkage

Rakshit Trivedi[1*], Bunyamin Sisman[2], Jun Ma[2], Christos Faloutsos[2,3], Hongyuan Zha[1], Xin Luna Dong[2]

[1]Georgia Institute of Technology, [2]Amazon.com, [3]Carnegie Mellon University

Multi-Graph Representation Learning

- Knowledge Graphs – Model facts using triplet form (e^s, r, e^o)
- Important resource for reasoning in NLP, QA systems, Dialogue Systems, Search, Recommendation
- Freebase (acquired by Google), Satori (Microsoft), Product Graph (Amazon), WikiData, DBpedia

Problem Definition

- How to learn entity and relation representations across multiple incomplete KGs: X, Y?

Existing Approach - Graph Alignment → Representation Learning

LinkNBed Method - Jointly Learn Representations and Entity Linkage

Key Insights

- ***Insight 1 (Embedding Similarity)***: If the two entities $e^X \in X$ and $e^Y \in Y$ represent the same real-world entity then their embeddings $\mathbf{e}^{\mathbf{X}}$ and $\mathbf{e}^{\mathbf{Y}}$ will be close to each other
- ***Insight 2 (Semantic Replacement)***: For a given triplet $t = (e^s, r, e^o) \in X$, denote $g(t)$ as the function that computes a relational score for t. If there exists a matching entity $e^{s'} \in Y$ for $e^s \in X$, denote $t' = (e^{s'}, r, e^o)$ obtained after replacing e^s with $e^{s'}$. In this case, $g(t) \sim g(t')$ i.e. score of triplets t and t' will be similar

LinkNBed Components

Atomic Layer

$$\mathbf{v}^{\mathbf{e^s}} = f(\mathbf{W}^{\mathbf{E}}\mathbf{e}^{\mathbf{s}}) \qquad \mathbf{v}^{\mathbf{e^o}} = f(\mathbf{W}^{\mathbf{E}}\mathbf{e}^{\mathbf{o}}) \qquad \mathbf{a} = \underbrace{f(\mathbf{W}^{\mathbf{key}}\mathbf{a}_{\mathbf{key}} + \mathbf{W}^{\mathbf{val}}\mathbf{a}_{\mathbf{val}})}_{\text{Attribute Network}}$$

$$\mathbf{v}^{\mathbf{r}} = f(\mathbf{W}^{\mathbf{R}}\mathbf{r}) \qquad \mathbf{v}^{\mathbf{t}} = f(\mathbf{W}^{\mathbf{T}}\mathbf{t})$$

Contextual Layer

$$\mathbf{c}(z) = \mathrm{AGG}(\mathbf{q}(z) * \{\mathbf{z}', \forall z' \in C(z)\}) \qquad \mathbf{q}(z) = \frac{\exp(\theta_z)}{\sum_{z' \in C(z)} \exp(\theta_{z'})}$$

- Entity Neighborhood Context - AGG: Mean, C: Rand. Walk
- Atrribute Context - AGG: Max, C: All attributes of an entity
- Type Context - AGG: Mean, C: All types of entities connected to a relation

Representation Layer

$$\mathbf{z}^{\mathbf{e^s}} = \sigma(\underbrace{\mathbf{W_1 v^{e^s}}}_{\text{Subject Entity Embedding}} + \underbrace{\mathbf{W_2 N_c}(e^s)}_{\text{Neighborhood Context}} + \underbrace{\mathbf{W_3 A_c}(e^s))}_{\text{Subject Entity Attributes}} \quad (1)$$

$$\mathbf{z}^{\mathbf{e^o}} = \sigma(\underbrace{\mathbf{W_1 v^{e^o}}}_{\text{Object Entity Embedding}} + \underbrace{\mathbf{W_2 N_c}(e^o)}_{\text{Neighborhood Context}} + \underbrace{\mathbf{W_3 A_c}(e^o))}_{\text{Object Entity Attributes}} \quad (2)$$

$$\mathbf{z}^{\mathbf{r}} = \sigma(\underbrace{\mathbf{W_4 v^r}}_{\text{Relation Embedding}} + \underbrace{\mathbf{W_5 T_c}(r))}_{\text{Entity Type Context}} \quad (3)$$

Relational Score Function

$$g(e^s, r, e^o) = \sigma(\mathbf{z}^{\mathbf{r}^T} \cdot (\mathbf{z}^{\mathbf{e^s}} \odot \mathbf{z}^{\mathbf{e^o}})) \quad (4)$$

LinkNBed Architecture

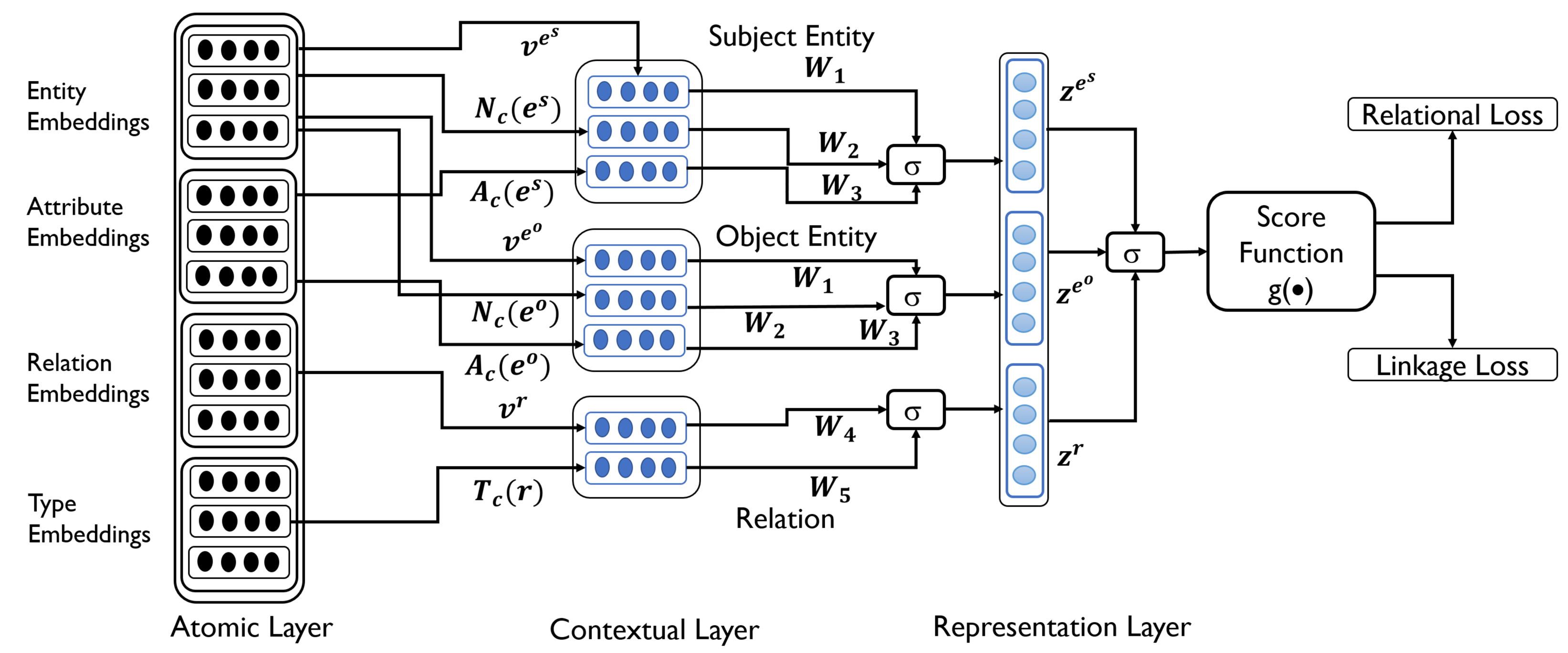

Figure: One step score computation for a given triplet (e^s, r, e^o). Attribute embeddings are learned based on attribute network defined in Atomic Layer Equation

Multi-Task Objective Function

Task 1: Relational Learning Loss

$$L_{rel} = \sum_{c=1}^{C} \max(0, \gamma - g(e_p^s, r_p, e_p^o) + g'(e_c^s, r_p, e_p^o)) \quad (5)$$

Task 2: Label Learning Loss

$$L_{lab} = \sum_{z=1}^{Z} \max(0, \gamma - g(e_Y^+, r_X, e_X^o) + (g'(e_Y^-, r_X, e_X^o)_z)) \quad (6)$$

Applying Multi-task Learning

$$\mathbf{L}(\psi) = \sum_{i=1}^{N} [b \cdot L_{rel} + (1-b) \cdot L_{lab}] + \lambda \|\psi\|_2^2 \quad (7)$$

- Parameter Space: $\psi = \{\{\mathbf{W_i}\}_{i=1}^5, \mathbf{W^E}, \mathbf{W^R}, \mathbf{W^{key}}, \mathbf{W^{val}}, \mathbf{W^t}, \theta\}$
- Training: Mini-batch SGD with Adam Optimizer
- Supports Missing Positive Labels and No Labels (Unsupervised)

Evaluation Protocol

Link Prediction

- Standard KG link prediction protocol used to report HITS@10 (predicted rank ≤ 10) and reciprocal rank ($\frac{1}{rank}$) over all test samples

Entity Linkage

- Novel evaluation scheme based on Insight 2.
- Algorithm 2 computes linkage score q ($\in [0, 1]$) for the pair of entities (e_X, e_Y).
- We use q to compute AUPRC for performance evaluation

Algorithm 2 Entity Linkage Score Computation

Input: Test pair – ($e_X \in X, e_Y \in Y$).
Output: Linkage Score – q.

1. Collect all triplets involving e_X from graph X and all triplets involving e_Y from graph Y into a combined set $\mathcal{O}$. Let $|\mathcal{O}| = k$.
2. Construct $S_{orig} \in \mathbb{R}^k$.
For each triplet $o \in \mathcal{O}$, compute score $g(o)$ using Eq. 10 and store the score in S_{orig}.
3. Create triplet set $\mathcal{O}'$ as following:
if $o \in \mathcal{O}$ contain $e^X \in X$ **then**
  Replace e^X with e^Y to create perturbed triplet o' and store it in $\mathcal{O}'$
end if
if $o \in \mathcal{O}$ contain $e^Y \in Y$ **then**
  Replace e^Y with e^X to create perturbed triplet o' and store it in $\mathcal{O}'$
end if
4. Construct $S_{repl} \in \mathbb{R}^k$.
For each triplet $o' \in \mathcal{O}'$, compute score $g(o')$ using Eq. 10 and store the score in S_{repl}.
5. Compute q.
Elements in S_{orig} and S_{repl} have one-one correspondence so take the mean absolute difference:
$q = |S_{orig} - S_{repl}|_1$
return q

Experimental Setup

Dataset Name	# Entities	# Relations	# Attributes	# Entity Types	# Available Triples
D-IMDB	378,207	38	23	41	143,928,582
D-FB	39,667	146	69	324	22,140,475

Table: Dataset Statistics: D-IMDB and D-FB (derived from IMDB and Freebase resp.)

Existing Relational Baselines: RESCAL, DISTMULT, Complex, STransE, GAKE
LinkNBed Variants: Embed Only, Attr Only, Nhbr Only, Embed + Attr, Embed + Nhbr, Embed All

Predictive Analysis

Link Prediction Results

Method	D-IMDB-HITS10	D-IMDB-MRR	D-FB-HITS10	D-FB-MRR
RESCAL	75.3	0.592	69.99	0.147
DISTMULT	79.5	0.691	72.34	0.556
Complex	83.2	0.725	75.67	0.629
STransE	80.7	0.421	69.87	0.397
GAKE	69.5	0.114	63.22	0.093
LinkNBed-Embed Only	79.9	0.612	73.2	0.519
LinkNBed-Attr Only	82.2	0.676	74.7	0.588
LinkNBed-Nhbr Only	80.1	0.577	73.4	0.572
LinkNBed-Embed + Attr	84.2	0.673	78.39	0.606
LinkNBed-Embed + Nhbr	81.7	0.544	73.45	0.563
LinkNBed-Embed All	84.3	0.725	80.2	0.632
LinkNBed-Embed All (Attention)	**86.8**	**0.733**	**81.9**	**0.677**
Improvement (%)	**4.15**	**1.10**	**7.61**	**7.09**

Table: HITS@10 and MRR performance for Link Prediction

Entity Linkage Results

Method	AUPRC (Supervised)	AUPRC (Unsupervised)
RESCAL	-	0.327
DISTMULT	-	0.292
Complex	-	0.359
STransE	-	0.231
GAKE	-	**0.457**
LinkNBed-Embed Only	0.376	0.304
LinkNBed-Attr Only	0.451	0.397
LinkNBed-Nhbr Only	0.388	0.322
LinkNBed-Embed + Attr	0.512	0.414
LinkNBed-Embed + Nhbr	0.429	0.356
LinkNBed-Embed All	0.686	0.512
LinkNBed-Embed All (Attention)	**0.691**	**0.553**
Improvement (%)	**33.86**	**17.35**

Table: AUPRC for Entity Linkage - Unsupervised case uses MLP classifier at 2nd step

Exploratory Remarks

- Attributes have higher impact than other contexts on overall performance
- More sophisticated attention mechanism could help capture better discriminative features
- Aligning relationships, types and attributes in addition to entities may provide further improvement

Email correspondence: rstrivedi@gatech.edu