# Improving Knowledge Graph Embedding Using Simple Constraints

ACL 2018

BoyangDing, QuanWang, BinWang, LiGuo

Xiaofan Yan

# Outline

- Abstract
- Introduction
- Related Work
- Approach
- Experiments and Results
- Conclusion

# Abstract

- Embedding KG into continuous vector spaces
  - Early: KG triples
  - Recent:
    - Complicated triple scoring
    - Incorporating other information beyoud triples
- **Investigate --** constraints
  - Entity -- non-negative -- compact & interpretable
  - Relation -- approximate entailment -- regularities of logical entailment

# Introduction

- 🔥 embed components of a KG into a continuous vector space

  -- to simple manipulation while preserving the inherent structure of the KG

# Related Work

？： Grounding

- very simple models:
  - TransE
  - RESCAL
- design more complicated triple scoring models:
  - the TransE extensions
  - the RESCAL extensions
  - the(deep)neural network models
  - tried to integrate extra information beyond triples
    - entity types, relation paths, textual descriptions

- **Difference**
  - impose constraints directly on entity and relation representations **without grounding**.
  - Use are quite universal, requiring **no manual effort** and applicable to almost all KGs.

# Approach

？： $x = \mathrm{Re}(x) + i\mathrm{In}(x)$

- A Basic Embedding Model -- ComplEx(Trouillonetal.,2016)
- A KG containing a set of triples $\mathcal{O} = \{(e_i, r_k, e_j)\}$
- Score of the triple $\phi(e_i, r_k, e_j) \triangleq \mathrm{Re}(\langle \mathbf{e}_i, \mathbf{r}_k, \bar{\mathbf{e}}_j \rangle)$

$$\triangleq \mathrm{Re}(\sum_\ell [\mathbf{e}_i]_\ell [\mathbf{r}_k]_\ell [\bar{\mathbf{e}}_j]_\ell),$$

- asymmetry： $\phi(e_i, r_k, e_j) \neq \phi(e_j, r_k, e_i)$

# Approach

- Non-negativity of Entity Representations
- Background:
  - the distributed representations can be taken as feature vectors for entities
  - it would be uneconomicaltostoreallnegativepropertiesofan entity or a concept
  - As shown by Lee and Seung (1999), non-negativity, in most cases, will further induce sparsity and interpretability.

$$\mathbf{0} \leq \mathrm{Re}(\mathbf{e}), \mathrm{Im}(\mathbf{e}) \leq \mathbf{1}, \quad \forall e \in \mathcal{E}$$

# Approach

？：极限推导

- Approximate Entailment for Relations
  - Means an ordered pair of relations that the former approcimately entails the latter
  - *strict entailment

$$\phi(e_i, r_p, e_j) \leq \phi(e_i, r_q, e_j), \quad \forall e_i, e_j \in \mathcal{E}$$

$$\mathrm{Re}(\mathbf{r}_p) \leq \mathrm{Re}(\mathbf{r}_q), \quad \mathrm{Im}(\mathbf{r}_p) = \mathrm{Im}(\mathbf{r}_q)$$

$$\lambda\big(\mathrm{Re}(\mathbf{r}_p) - \mathrm{Re}(\mathbf{r}_q)\big) \leq \boldsymbol{\alpha},$$

$$\lambda\big(\mathrm{Im}(\mathbf{r}_p) - \mathrm{Im}(\mathbf{r}_q)\big)^2 \leq \boldsymbol{\beta}.$$

# Approach

？：a negative triple can be generated by randomly corrupting the head or the tail entity of a positive triple,

- The Overall Model

$$\min_{\Theta,\{\alpha,\beta\}} \boxed{\sum_{\mathcal{D}^+\cup\mathcal{D}^-} \log\left(1+\exp(-y_{ijk}\phi(e_i, r_k, e_j))\right)} + \boxed{\mu \sum\nolimits_{\mathcal{T}} \mathbf{1}^\top(\boldsymbol{\alpha}+\boldsymbol{\beta})} + \boxed{\eta\|\Theta\|_2^2}$$

**Logistic Loss** Enforces triples to have scores close to their labels.

**Sum of slack variables** hope the slackness to be small

$\mathbf{L_2}$ avoid over-fitting

$$\text{s.t.}\ \ \lambda\big(\mathrm{Re}(\mathbf{r}_p)-\mathrm{Re}(\mathbf{r}_q)\big)\le\boldsymbol{\alpha},\quad \boldsymbol{\alpha},\boldsymbol{\beta}\ge\mathbf{0},\quad \forall r_p\xrightarrow{\lambda} r_q\in\mathcal{T},$$
$$\lambda\big(\mathrm{Im}(\mathbf{r}_p)-\mathrm{Im}(\mathbf{r}_q)\big)^2\le\boldsymbol{\beta},\quad \mathbf{0}\le\mathrm{Re}(\mathbf{e}),\mathrm{Im}(\mathbf{e})\le\mathbf{1},\quad \forall e\in\mathcal{E}.$$

# Approach

- The Overall Model

$$\min_{\Theta,\{\alpha,\beta\}} \sum_{\mathcal{D}^+\cup\mathcal{D}^-} \log\left(1+\exp(-y_{ijk}\phi(e_i,r_k,e_j))\right) + \mu\sum\nolimits_{\mathcal{T}} \mathbf{1}^\top(\boldsymbol{\alpha}+\boldsymbol{\beta}) + \eta\|\Theta\|_2^2$$

$$\text{s.t.}\;\; \lambda\big(\mathrm{Re}(\mathbf{r}_p)-\mathrm{Re}(\mathbf{r}_q)\big) \le \boldsymbol{\alpha}, \quad \boldsymbol{\alpha},\boldsymbol{\beta} \ge \mathbf{0}, \quad \forall r_p \xrightarrow{\lambda} r_q \in \mathcal{T},$$

$$\lambda\big(\mathrm{Im}(\mathbf{r}_p)-\mathrm{Im}(\mathbf{r}_q)\big)^2 \le \boldsymbol{\beta}, \quad \mathbf{0} \le \mathrm{Re}(\mathbf{e}),\mathrm{Im}(\mathbf{e}) \le \mathbf{1}, \quad \forall e \in \mathcal{E}.$$

$$\min_{\Theta} \sum_{\mathcal{D}^+\cup\mathcal{D}^-} \log\left(1+\exp(-y_{ijk}\phi(e_i,r_k,e_j))\right) + \mu\sum\nolimits_{\mathcal{T}} \lambda\mathbf{1}^\top\big[\mathrm{Re}(\mathbf{r}_p)-\mathrm{Re}(\mathbf{r}_q)\big]_+$$

$$+\mu\sum\nolimits_{\mathcal{T}} \lambda\mathbf{1}^\top\big(\mathrm{Im}(\mathbf{r}_p)-\mathrm{Im}(\mathbf{r}_q)\big)^2 + \eta\|\Theta\|_2^2$$

$$\text{s.t.}\;\; \mathbf{0} \le \mathrm{Re}(\mathbf{e}),\mathrm{Im}(\mathbf{e}) \le \mathbf{1}, \quad \forall e \in \mathcal{E}$$

# Approach

- The Overall Model
- **optimizer:** SGD in mini-batch mode
- **to tune the learning rate:** AdaGrad (Duchi et al., 2011)
- project (by truncation) Re(x)&Im(x) **into the hypercube of $[0,1]^d$**

- **Time complexity:** $O(sd+td+\bar{n}d)$
- **Space complexity:** $O(nd+md)$

The same as **ComplEx**

# Experiments and Results

- **Datasets:**
  - WN18 + FB15K + DB100K -- divided into **training, validation, and test**
  - Use AMIE+ (Gal´arragaetal.,2015) to extract approximate entailments automatically from the training set of each dataset.
  - Consider entailments with PCAconfidencehigherthan0.8

| Dataset | # Ent | # Rel | # Train/Valid/Test | | | # Cons |
|---|---|---|---|---|---|---|
| WN18 | 40,943 | 18 | 141,442 | 5,000 | 5,000 | 17 |
| FB15K | 14,951 | 1,345 | 483,142 | 50,000 | 59,071 | 535 |
| DB100K | 99,604 | 470 | 597,572 | 50,000 | 50,000 | 56 |

# Experiments and Results

- **Link Prediction**
  - aims to predict a triple $(e_i,r_k,e_j)$ with $e_i$ or $e_j$ missing, i.e., predict $e_i$ given $(r_k,e_j)$ or predict $e_j$ given $(e_i,r_k)$.
  - **Evaluation Protocol:** (Bordes et al., 2013)
    - report on the test set the mean reciprocal rank (MRR) and the proportion of correct entities ranked in the top n (HITS@N), with n = 1,3,10.
  - **Comparison Settings:**
    - Simple embedding models that utilize triples alone without integrating extra information
    - Other extensions of ComplEx that integrate logical background knowledge in addition to triples
    - Latest developments or implementations that achieve current state-of-the-art performance
    - ComplEx-NNE
    - ComplEx-NNE+AER

# Experiments and Results

- **Implementation Details**

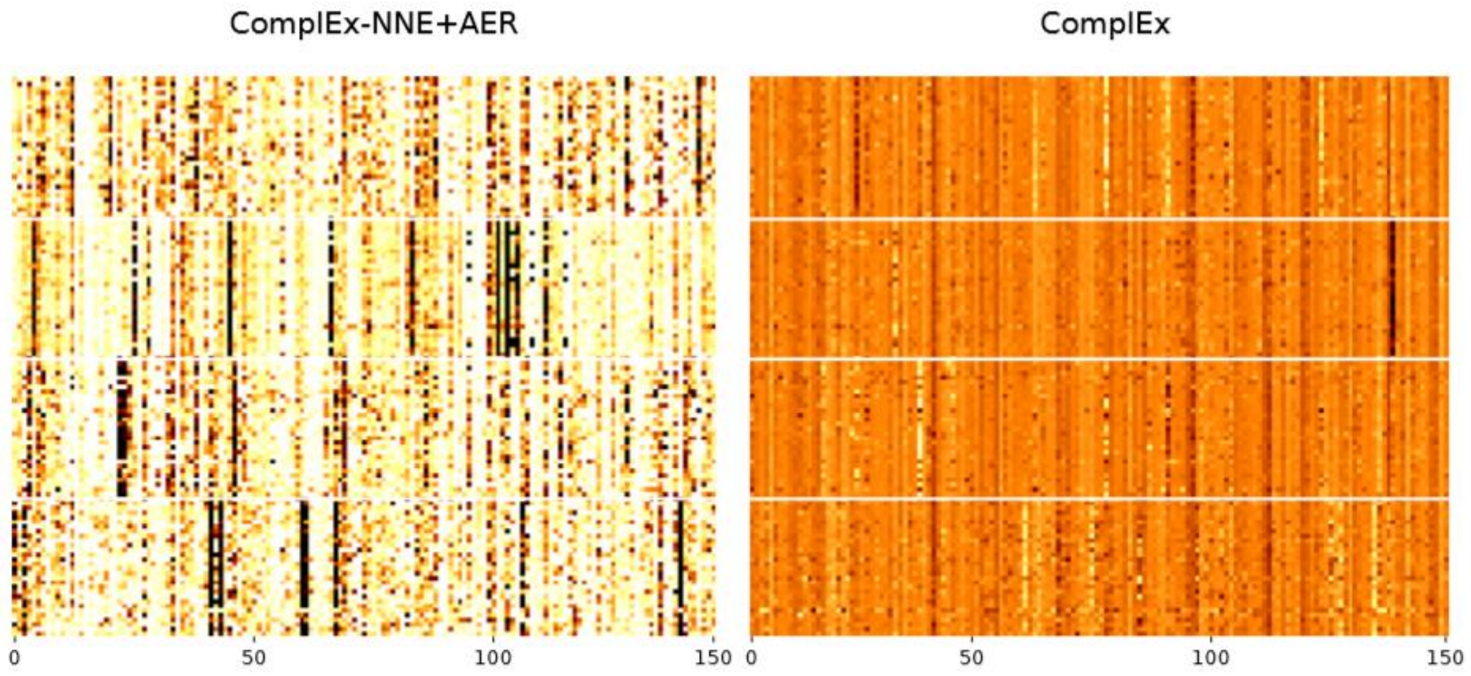
| WN18 | FB15K | DB100K |
|---|---|---|
| ✓ | ✓ | ✓ |
| ✓ | ✓ | ✓ |
| ✓ | ✓ | ✗ |
| ✓ | ✓ | ✓ |
| ✓ | ✓ | ✓ |

- Simple embedding models
- Other extensions of ComplEx
- Latest developments or implementations

- ComplEx-NNE
- ComplEx-NNE+AER

# Experiments and Results

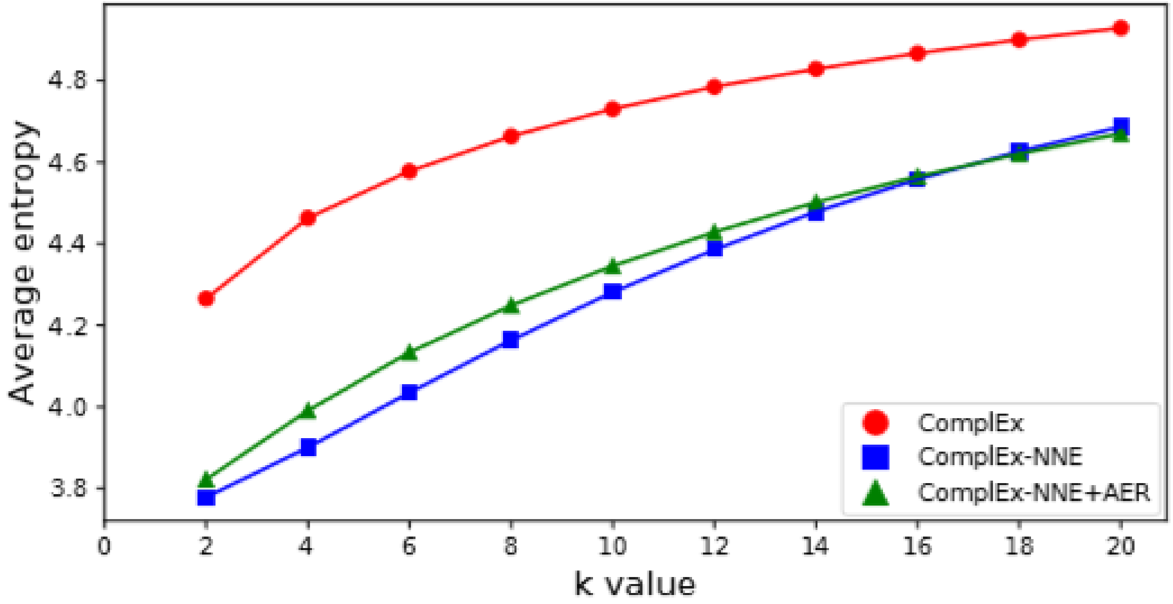
- **Experimental Results**




- **Link prediction：** compact and interpretation representation
  - From top to bottom, entities :reptile, wine region, species, and programming language.
  - Values range from 0 (white) via 0.5 (orange) to 1 (black). Best viewed in color.

# Experiments and Results

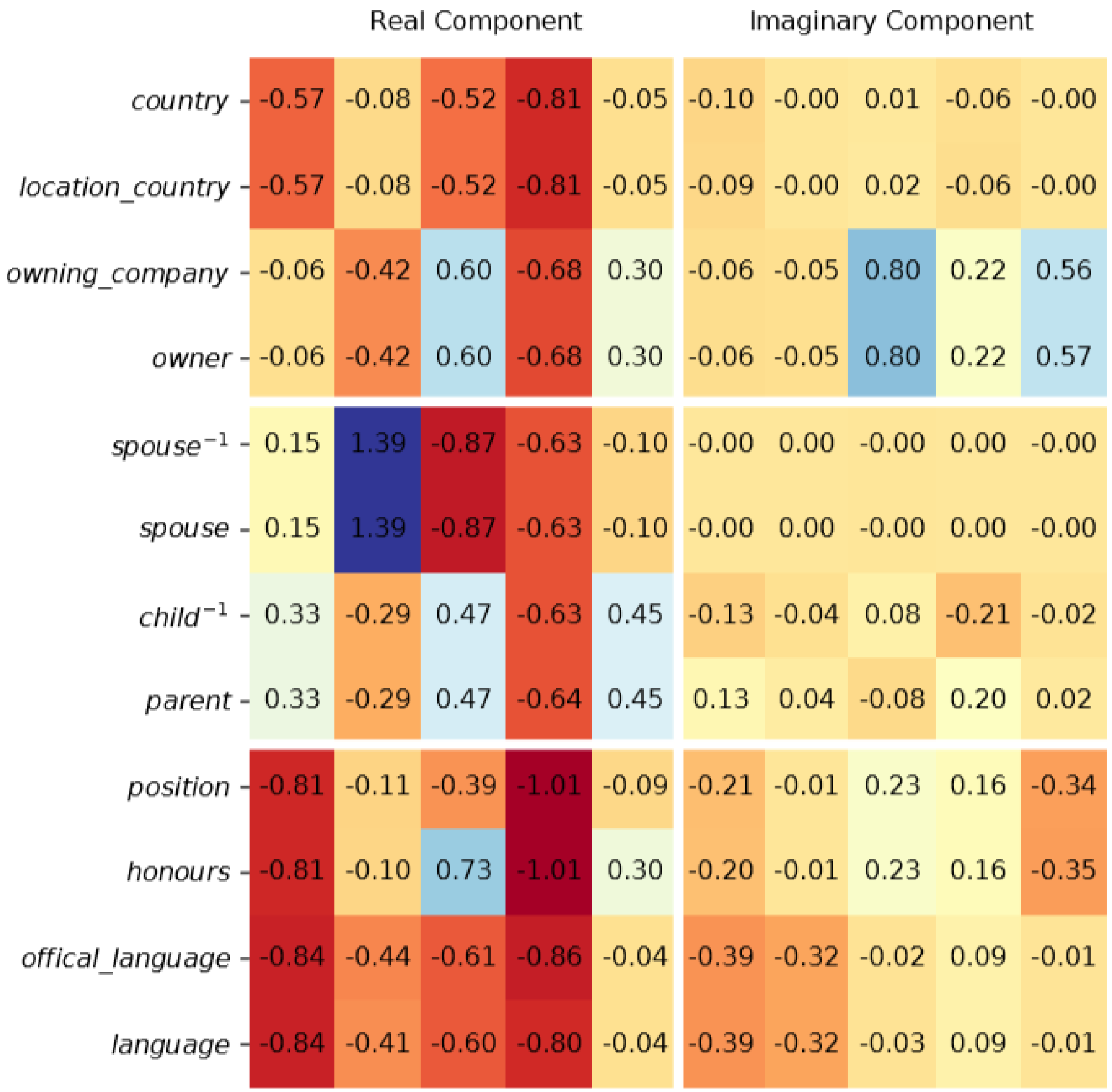
- **Experimental Results**




- **Average entropy:** over all dimensions of Re(x)
  - For each dimension, top K percent of entities with the highest activation values are picked. We can calculate the entropy of the type distribution of the entities selected. This entropy reflects **diversity of entity types**, or in other words, **semantic purity**.
  - ComplEx-NNE and ComplEx-NNE+AER can learn entity representations with latent dimensions of consistently higher semantic purity.

# Experiments and Results

- **Experimental Results**
- Top 4 relations from the equivalence class, the middle 4 the inversion class, and the bottom 4 others.
- **Encode logical regularities quite well**
  - Pairs of relations from the first class (equivalence) tend to have identical representations $r_p \approx r_q$
  - those from the second class (inversion) complex conjugate representations $r_p \approx \bar{r}_q$
  - and the others representations that $\mathrm{Re}(r_p) \leq \mathrm{Re}(r_q)$ and $\mathrm{Im}(r_p) \approx \mathrm{Im}(r_q)$



# Conclusion

- Two types of constraints
  - non-negativity constraints -- compact, interpretable entity representations
  - approximate entailment constraints -- encode logical regularities into relation representations
- Experimental results -- simple yet surprisingly effective
- The constraints indeed improve model interpretability, yielding a substantially increased structuring of the embedding space.