Investigating Robustness and Interpretability of Link Prediction via Adversarial Modifications

Pouya Pezeshkpour  
Yifan Tian  
Sameer Singh

---

**Graph Embeddings for Link Prediction**

In this work, we propose efficient adversarial modifications for link prediction models to evaluate robustness, and study interpretability and error correcting.

**Completing Knowledge Graphs:**
Predicting a missing link from observed graph structure.

- Existing models:
  - Embed s, r, and o
  - Maximize score \( \psi(s, r, o) \) for observed facts
    
    - DistMult: \( e_s R e_o \)
    
    - ConvE: \( f(\text{vec}(f((e_s; r; e_o))W)e_o) \)

- Embeddings are inscrutable...
  - Are these embeddings robust to small changes?
  - Can we explain why a fact/link was predicted?

**Adversarial Modifications (CRIAGE)**

- Completion Robustness and Interpretability via Adversarial Graph Edits (CRIAGE)

  Minimally change the graph so that target fact prediction changes the most after embeddings are relearned.

---

**Efficiently Identifying the Modification**

For target triple \( s, r, o \) and graph G, we identify:

- **Removing / Adding**: Find \( (s', r', o) \) such that score \( \psi(s, r, o) \) trained on \( G \) is maximally different from score \( \psi(s', r', o) \) trained after removing or adding \( (s', r', o) \):

\[
\arg\max_{(s', r', o)} \psi(s', r', o) - \psi(s, r, o)
\]

**Two Primary Challenges:**

1. **Retraining is too expensive**: Taylor approximation on gradient of loss and utilizing graph structure.

\[
\Delta(\nabla_{\text{loss}}(e)) = H_e(\text{loss}) \times (e - \bar{e})
\]

\( e, \bar{e} = \text{optimal embedding} \land H = \text{Hessian} \)

\[
\Rightarrow \bar{e} = e - H_e(\text{loss})^{-1} \times \Delta(\nabla_{\text{loss}}(e))
\]

2. **Too many links to search**: Learn a continuous space of links using an inverter, and use gradient descent.

---

**CRIAGE vs Influence Functions**

- **Influence Functions (IF)**:
  - Similar motivation, but doesn’t exploit graph structure

---

**Robustness and Interpretability**

**Robustness**

Does adding a fake link affect performance?

- Yago3 Hits@1 (Adding a fake link)
- WordNet Hits@1 (Adding a fake link)

**Interpretability**

Which link, when removed, changes the prediction?

Find common patterns in removed link \( R(a, b) \)

- DistMult and ConvE:
  - isMarriedTo(a,c) \land hasChild(c,b) \Rightarrow hasChild(a,b)
  - playFor(a,c) \land isLocatedIn(c,b) \Rightarrow wasBornIn(a,b)
  - isLocatedIn(c,b) \Rightarrow diedIn(a,b)
  - hasAdvisor(a,c) \land graduatedFrom(c,b) \Rightarrow graduatedFrom(a,b)
  - influences(a,c) \land influences(c,b) \Rightarrow influences(a,b)

* Identified as rules by [Yang et al. 2015]

**Error Correcting**

Introduce errors and see if we can detect it.

Choose neighbor \( w \) at least \( \psi(s, r, o) - \tilde{\psi}(s, r, o) \) as incorrect.

---

For any questions, email: pezeshkp@uci.edu

Website: https://pouyapez.github.io/criage